# THE JOURNAL OF PHILOSOPHY

Cuk - Ho5964-7-&P836

CALCUTTA UNIVERSITY LIBRARY
ALIPUR
CAMPUS

3-1-2000

VOLUME XCVI, NUMBER 2
FEBRUARY 1999

| page | | |
|---|---|---|
| 55 | *Second-hand Moral Knowledge* | Karen Jones |
| 79 | *Restitution and Revenge* | David B. Hershenov |
| | BOOK REVIEWS | |
| 95 | Alfred R. Mele: *Autonomous Agents: From Self-control to Autonomy* | John Christman |
| 99 | Noël Carroll: *A Philosophy of Mass Art* | Gregg M. Horowitz |
| 106 | NEW BOOKS: ANTHOLOGIES | |

*Published by* **The Journal of Philosophy, Inc.**

# THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

*Editors*: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Achille C. Varzi. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

# THE JOURNAL OF PHILOSOPHY

1999

SUBSCRIPTIONS (12 issues)

| | |
|---|---|
| Individuals | $35.00 |
| Libraries and Institutions | $75 00 |
| Students, retired/unemployed philosophers | $20.00 |
| Postage outside the U.S. | $15.00 |

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1999 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

# THE JOURNAL OF PHILOSOPHY

CALCUTTA UNIVERSITY LIBRARY
ALIPUR
CAMPUS

3-1-2000

VOLUME XCVI, NUMBER 1
JANUARY 1999

7

page
5 *Atomic Metaphysics* Nick Huggett
25 *Seeing Intersecting Eclipses* Roy Sorensen

BOOK REVIEWS
50 George Boolos: *The Logic of Provability* W. W. Tait

54 MEMO TO AUTHORS

*Published by* The Journal of Philosophy, Inc.

# THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

*Editors*: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Achille C. Varzi. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

KP-836 (20.4.2001).

# THE JOURNAL OF PHILOSOPHY

1999

SUBSCRIPTIONS (12 issues)

| | |
|---|---|
| Individuals | $35.00 |
| Libraries and Institutions | $75.00 |
| Students, retired/unemployed philosophers | $20.00 |
| Postage outside the U.S. | $15.00 |

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742.

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1999 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

# THE JOURNAL OF
# PHILOSOPHY

3-1-2000

CALCUTTA UNIVERSITY LIBRARY
ALIPUR
CAMPUS

VOLUME XCVI, NUMBER 3
MARCH 1999

page
109 *Of Conspiracy Theories* Brian L. Keeley
127 *The Emperor's New Intuitions* Jaakko Hintikka

COMMENTS AND CRITICISM
148 *Supervaluation Can Leave Truth-value Gaps after All*
Michael Morreau

157 NEW BOOKS: ANTHOLOGIES
162 NOTES AND NEWS

***Published by*** **The Journal of Philosophy, Inc.**

# THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

*Editors*: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Achille C. Varzi. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

# THE JOURNAL OF PHILOSOPHY

1999

SUBSCRIPTIONS (12 issues)

| | |
|---|---|
| Individuals | $35.00 |
| Libraries and Institutions | $75.00 |
| Students, retired/unemployed philosophers | $20.00 |
| Postage outside the U.S. | $15.00 |

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1999 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

# THE JOURNAL OF PHILOSOPHY

8-1-2000

CALCUTTA UNIVERSITY LIBRARY
ALIPUR
CAMPUS

VOLUME XCVI, NUMBER 4
APRIL 1999

page
163 *Is Two a Property?* Byeong-uk Yi

COMMENTS AND CRITICISM
191 *Wings, Spoons, Pills, and Quills: A Pluralist Theory of Function*
Ruth Garrett Millikan

BOOK REVIEWS
207 Jerrold J. Katz: *Realistic Rationalism* Michael D. Resnik

211 NEW BOOKS

*Published by* **The Journal of Philosophy, Inc.**

# THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

*Editors*: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Achille C. Varzi.
*Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

# THE JOURNAL OF PHILOSOPHY

1999

SUBSCRIPTIONS (12 issues)

| | |
|---|---|
| Individuals | $35.00 |
| Libraries and Institutions | $75.00 |
| Students, retired/unemployed philosophers | $20.00 |
| Postage outside the U.S. | $15.00 |

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742.

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1999 by the Journal of Philosophy, Inc.

ISSN 0022-362X
POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

# THE JOURNAL OF PHILOSOPHY

CALCUTTA UNIVERSITY LIBRARY
ALIPUR
CAMPUS

3-1-2000

VOLUME XCVI, NUMBER 5
MAY 1999

page
217 *Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism* Robert Kane
241 *Intention and Weakness of Will* Richard Holton

BOOK REVIEWS
263 Jonathan Lear: *Open Minded: Working Out the Logic of the Soul* Marcia Cavell

270 NEW BOOKS

***Published by*** **The Journal of Philosophy, Inc.**

# THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

*Editors*: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Achille C. Varzi.
*Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX. (212) 932-3721.

# THE JOURNAL OF PHILOSOPHY

1999

SUBSCRIPTIONS (12 issues)

| | |
|---|---|
| Individuals | $35.00 |
| Libraries and Institutions | $75.00 |
| Students, retired/unemployed philosophers | $20.00 |
| Postage outside the U.S. | $15.00 |

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742.

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1999 by the Journal of Philosophy, Inc. ISSN 0022-362X
POSTMASTER Periodical postage paid at New York, NY, and other mailing offices.

# THE JOURNAL OF PHILOSOPHY

3-1-2000

CALCUTTA UNIVERSITY LIBRARY
ALIPUR
CAMPUS

VOLUME XCVI, NUMBER 6
JUNE 1999

*page*
271 *Internalism Exposed* Alvin I. Goldman
294 *Calibration and the Epistemological Role of Bayesian Conditionalization* Marc Lange

**_Published by_ The Journal of Philosophy, Inc.**

# THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

*Editors*: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Achille C. Varzi.
*Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX. (212) 932-3721.

# THE JOURNAL OF PHILOSOPHY

1999

SUBSCRIPTIONS (12 issues)

| | |
|---|---|
| Individuals | $35.00 |
| Libraries and Institutions | $75.00 |
| Students, retired/unemployed philosophers | $20.00 |
| Postage outside the U.S. | $15.00 |

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742.

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1999 by the Journal of Philosophy, Inc.

ISSN 0022-362X
POSTMASTER Periodical postage paid at New York, NY, and other mailing offices

# THE JOURNAL OF
# PHILOSOPHY

CALCUTTA UNIVERSITY LIBRARY
ALIPUR
CAMPUS

3-1-2000

VOLUME XCVI, NUMBER 7
JULY 1999

*page*
325 *Presentism and Ontological Commitment* Theodore Sider
348 *Reducing Thermodynamics to Statistical Mechanics: The Case of Entropy* Craig Callender
374 NEW BOOKS

*Published by* **The Journal of Philosophy, Inc.**

# THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

*Editors*: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Achille C. Varzi. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

# THE JOURNAL OF PHILOSOPHY

1999

SUBSCRIPTIONS (12 issues)

| | |
|---|---|
| Individuals | $35.00 |
| Libraries and Institutions | $75.00 |
| Students, retired/unemployed philosophers | $20.00 |
| Postage outside the U.S. | $15.00 |

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742.

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1999 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

# THE JOURNAL OF PHILOSOPHY

CALCUTTA UNIVERSITY LIBRARY
ALIPUR
CAMPUS

3-1-2000

VOLUME XCVI, NUMBER 8
AUGUST 1999

*page*
381 *All at Sea in Semantic Space: Churchland on Meaning Similarity* Jerry Fodor and Ernie Lepore
404 *On a Supposed Counterexample to Modus Ponens* Bernard D. Katz

COMMENTS AND CRITICISM
416 *Semantics San Diego Style* Evan Tiffany

BOOK REVIEWS
430 John Haugeland: *Having Thought: Essays in the Metaphysics of Mind* Daniel C. Dennett

436 NEW BOOKS

***Published by*** **The Journal of Philosophy, Inc.**

# THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

*Editors*: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Achille C. Varzi. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

# THE JOURNAL OF PHILOSOPHY

1999

SUBSCRIPTIONS (12 issues)

| | |
|---|---|
| Individuals | $35.00 |
| Libraries and Institutions | $75.00 |
| Students, retired/unemployed philosophers | $20.00 |
| Postage outside the U.S. | $15.00 |

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1999 by the Journal of Philosophy, Inc. ISSN 0022-362X
POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices

# THE JOURNAL OF PHILOSOPHY

CALCUTTA UNIVERSITY LIBRARY
ALIPUR
CAMPUS

3-1-2000

VOLUME XCVI, NUMBER 9
SEPTEMBER 1999

*page*
437 *Measuring Confirmation* David Christensen
462 *Nothing, Something, Infinity* Joseph Almog
479 *Putnam's Model-theoretic Argument Reconstructed* Igor Douven

***Published by*** **The Journal of Philosophy, Inc.**

# THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

*Editors*: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Achille C. Varzi. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

# THE JOURNAL OF PHILOSOPHY

1999

SUBSCRIPTIONS (12 issues)

| | |
|---|---|
| Individuals | $35.00 |
| Libraries and Institutions | $75.00 |
| Students, retired/unemployed philosophers | $20.00 |
| Postage outside the U.S. | $15.00 |

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1999 by the Journal of Philosophy, Inc. ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

# THE JOURNAL OF PHILOSOPHY

VOLUME XCVI, NUMBER 10
OCTOBER 1999

page
491 *Is There a Place for Philosophy in Quine's Theory?* Gila Sher
525 *Can a Turing Machine Know that the Gödel Sentence Is True?* Storrs McCall

COMMENTS AND CRITICISM
533 *Deflating the Conservativeness Argument* Hartry Field
541 *Comments on Shapiro* Jody Azzouni

1-2-2000

CALCUTTA UNIVERSITY LIBRARY
ALIPUR
CAMPUS

*Published by* **The Journal of Philosophy, Inc.**

# THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

*Editors*: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Achille C. Varzi. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

# THE JOURNAL OF PHILOSOPHY

1999

SUBSCRIPTIONS (12 issues)

| | |
|---|---|
| Individuals | $35.00 |
| Libraries and Institutions | $75.00 |
| Students, retired/unemployed philosophers | $20.00 |
| Postage outside the U.S. | $15.00 |

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1999 by the Journal of Philosophy, Inc. ISSN 0022-362X
POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices

# THE JOURNAL OF
# PHILOSOPHY

VOLUME XCVI, NUMBER 11
NOVEMBER 1999

page
545 *Meaning and Deflationary Truth* Michael Williams
565 *Knowing Failably* Stephen Hetherington
588 *Bargaining with Neighbors: Is Justice Contagious?*
Jason Alexander and Brian Skyrms

CALCUTTA UNIVERSITY LIBRARY
ALIPUR
CAMPUS

***Published by*** **The Journal of Philosophy, Inc.**

# THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

*Editors*: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Achille C. Varzi. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721

# THE JOURNAL OF PHILOSOPHY

1999

SUBSCRIPTIONS (12 issues)

| | |
|---|---|
| Individuals | $35.00 |
| Libraries and Institutions | $75.00 |
| Students, retired/unemployed philosophers | $20.00 |
| Postage outside the U.S. | $15.00 |

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us

© Copyright 1999 by the Journal of Philosophy, Inc.

ISSN 0022-362X
POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

# THE JOURNAL OF PHILOSOPHY

VOLUME XCVI, NUMBER 12
DECEMBER 1999

page

599 *The Moral Basis of Political Liberalism* Charles Larmore

BOOK REVIEWS

626 John Searle: *Mind, Language, and Society: Philosophy in the Real World* Simon Blackburn
630 Anthony J. Cascardi: *Consequences of Enlightenment* Michael Kelly
638 Solomon Feferman: *In the Light of Logic* Jeremy Avigad

643 INDEX TO VOLUME XCVI

CALCUTTA UNIVERSITY LIBRARY
ALIPUR
CAMPUS

*Published by* **The Journal of Philosophy, Inc.**

# THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

*Editors*: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Achille C. Varzi. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

# THE JOURNAL OF PHILOSOPHY

1999

SUBSCRIPTIONS (12 issues)

| | |
|---|---|
| Individuals | $35.00 |
| Libraries and Institutions | $75.00 |
| Students, retired/unemployed philosophers | $20.00 |
| Postage outside the U.S. | $15.00 |

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1999 by the Journal of Philosophy, Inc.

ISSN 0022-362X
POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices

3-1-2000
CALCUTTA UNIVERSITY LIBRARY
ALIPUR
CAMPUS

# THE JOURNAL OF PHILOSOPHY

THE JOURNAL OF PHILOSOPHY was founded in 1904 as THE JOURNAL OF PHILOSOPHY, PSYCHOLOGY AND SCIENTIFIC METHODS by Frederick J. E. Woodbridge and J. McKeen Cattell. In 1906, Wendell T. Bush became associated with the JOURNAL as co-editor. The title was changed in 1921, and the JOURNAL was incorporated in the State of New York under its present name in 1923. Since its founding, it has been published from the Department of Philosophy at Columbia University. It is typeset and printed by Cadmus Journal Services.

# THE JOURNAL OF PHILOSOPHY

VOLUME XCVI, NUMBER 1
JANUARY–DECEMBER 1999

*Edited by*
Brian Barry, Bernard Berofsky
Akeel Bilgrami, Arthur C. Danto
Kent Greenawalt, Patricia Kitcher
Isaac Levi, Mary Mothersill
Philip Pettit

*Published by* The Journal of Philosophy, Inc.

# STATEMENT OF COPYRIGHT POLICY

We would like to draw your attention to the protection copyright affords against multiple photocopying by libraries, instructors, and copy centers. The 1976 copyright law provides that photocopies in excess of "fair use" copying cannot be made without authorization by the copyright owner—"fair use" meaning single copying for personal use. Accordingly, the Journal authorizes personal or educational multiple copying only if a request is made in writing, and the fee of $.05 per page/per copy is paid directly to us. *Each copy must include a notice of copyright.* We shall turn over to the authors any substantial payments (over $30.00, an administrative fee) obtained from all photocopying. Inquiries about photocopying may be sent to the Managing Editor.

© Copyright 1999 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER Second class postage paid at New York, NY, and other mailing offices. Please send address changes to the Journal of Philosophy, 709 Philosophy Hall, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, NY, NY 10027 Fax: (212) 932-3721

THE JOURNAL OF PHILOSOPHY

VOLUME XCVI, NO. 1, JANUARY 1999

# ATOMIC METAPHYSICS*

According to Dr Seuss, the father of the father of Nadd "knows everything in all [the] kingdom, in all the world beyond, and in all the other worlds that may happen to be."—*The Five Hundred Hats of Bartholomew Cubbins*

As a simple statistical model, consider a pair of coins repeatedly flipped. From experience, the usual frequency is heads-heads (or tails-tails) one time in four, and heads-tails one time in two. A second possibility is that typical in quantum mechanics; heads-heads, tails-tails, and heads-tails each occur one time in three. Often these frequencies are known as *Maxwell-Boltzmann* and *Bose-Einstein*, respectively. It is suggested—in particular by standard physics texts—that this difference is the important philosophical innovation of quantum mechanics for statistical physics. The aim of this paper is to construct a precise framework for judging this claim, one which reveals a considerable confusion in standard pronouncements. More precisely, I shall demonstrate that the occurrence of Maxwell-Boltzmann statistics in classical mechanics has none of the metaphysical implications that it is often thought to have. In so doing, I show that the terms 'Maxwell-Boltzmann' and 'Bose-Einstein' are themselves ambiguous, so I shall not use them until matters have been clarified.

If we suppose that these statistics are reflections of a uniform measure over universes of possible worlds—distinct physical states—then

---

* This paper is based on work submitted for a Ph.D. at Rutgers University. Earlier versions have been presented at the University of Manchester, Columbia University, the University of Kentucky, the University of Illinois/Chicago, the London School of Economics, and Leeds University. I wish to thank all concerned for their hospitality and helpful comments. Special thanks to Robert Weingard and Tim Maudlin.

0022-362X/99/9601/5–24

the two kinds have different implications for the universes to which they belong. The first, 1:2:1, implies that the modal universe contains two states that are heads-tails for every one that is heads-heads. Of course, this is not a surprising result in our world, for 'heads-or-tails' describes only a few of any coin's properties. There is only one way for both the quarter I got in my change at "K-Mart," and the one from "The Frog and the Peach" to be heads, but the "K-Mart" coin could be either heads or tails if they fall heads-tails. Things become interesting when we have statistics for a complete set of properties. For instance, consider the kind of qualitatively impoverished world in which the only positive, logically simple properties available to a coin are heads and tails. If there are now two states in which we find the result heads-tails, then there are two qualitatively indistinguishable worlds that cannot be identified. Does the difference not lie purely in the way that properties are distributed over individuals? Now this coin is heads, now that.

We appear to have an argument that takes us from an experiment involving coins to metaphysical conclusions about the structure of the modal universe of possible physical states. It seems that the observed statistics are to be explained by a mini physical theory that distinguishes four possibilities, including two worlds (the heads-tails worlds) that are qualitatively indistinguishable and yet distribute qualities differently to individuals. Loosely, this implies a denial of the Identity of Indiscernibles across possible worlds: a distinct coin is tails in each of the two worlds. This denial is the doctrine of *haecceitism*, which I shall clarify below. What motivates this discussion, then, is the suggestion that this significant metaphysical issue[1] can be settled experimentally.

The job of section I is to explicate the notion of haecceitism, and to distinguish it from various other aspects of individuality with which it is easily conflated. In subsequent sections (II and III) we shall generalize from the coins example, and see in a rigorous way what haecceitism means in physically realistic systems. Section IV will outline empirical arguments for and against haecceitistic representations in classical physics. Sections V and VI will demonstrate why they fail; after all, there is no experimental way to determine whether the classical world is haecceitistic or not. In the final section, I shall explore the consequences of this conclusion for the metaphysics of classical mechanics.

---

[1] Significant, for instance, to the theory of rigid designation: Saul Kripke, *Naming and Necessity* (Cambridge. Harvard, 1980), pp. 15-20, touches this point.

## I HAECCEITISM AND INDIVIDUALITY

To evaluate the truth values of modal propositions, we need to know what things in one world, say, ours, are getting up to in another: for instance, is there any world in which I present a television game show? However this goal is achieved, following David Lewis[2] we shall say of someone in our world, that another world *represents de re* of him that he has whatever properties "he" (perhaps his counterpart, or he himself) has in that world. So, worlds in which "I" am on the box represent de re of me that I am a television personality. Now we can precisely state the doctrine of haecceitism:

> If two worlds differ in what they represent *de re* concerning some individual, but do not differ qualitatively in any way, I shall call that a *haecceitistic difference. Haecceitism* is the doctrine that there are at least some cases of haecceitistic difference between worlds. *Anti-haecceitism* is the doctrine that there are not (*ibid.*, p. 221).[3]

Suppose we find conventional (one time in two) statistics for heads-tails, then (still supposing these to be the only properties that coins have) we might infer the existence of two worlds (both with one coin heads and one tails) that are qualitatively identical. But we take it that they do differ in their de re representations, for the heads coin in one world is tails in the other, and vice versa. Hence there is a haecceitistic difference between the worlds, and their modal universe is haecceitistic.

By way of further explication of this doctrine, recall what might motivate it in the first place. For instance, is it possible that 'I could have been John Kuti' is true? Not just in the sense that I could have gone to law school, that I could have had a friend called Nick Huggett, and that I could have moved to Moscow. Rather, could I have had the very same life that John has in this world? It seems that this kind of statement must be false on any account of my otherworldly selves which picks them out qualitatively. The person with John's life will always be picked out as John, not me. But if we have a nonqualitative way of picking ourselves out in other worlds, then we open up the possibility that we could get up to different things in qualitatively identical worlds. So, asserting the truth of such statements seems to drive one toward haecceitism.

Extreme caution is needed here, for it is very easy to slip from haecceit*ism* into considerations of haecceit*ies*; differences in de re representation in the face of qualitative identity could be explained by nonqualitative properties of "thisness" co-instantiated with qualities in different ways. In one case, for instance, the haecceity thisness-of-coin-1 might be co-instantiated with the atomic quality of being-heads, but in a qualitatively identical world, with being-tails.[4] But (as Lewis points out) haecceitism is neither necessary nor sufficient for haecceities. For instance, in two identical worlds of individuals that maximally violate the Identity of Indiscernibles—so that all the individuals in both worlds are in identical states—both worlds represent the same of any individual, and so haecceitism fails. But there could still be nonqualitative properties differentiating the individuals. On the other hand, differences in representation de re do not have to be explained in terms of haecceities, for they could be brute metaphysical facts.

---

[2] *On the Plurality of Worlds* (New York· Blackwell, 1986), § 4.4.

[3] For a discussion of the idea of "qualitative character," see *ibid.*, p. 22.

The point of this caution is to emphasize that we are investigating the well-defined notion of haecceitism, not the separate, murky notion of an haecceity. Our resolve will allow us to acknowledge that haecceitism is only one aspect of individuality and that other related concepts warrant separate investigation. In particular, we are interested in the converse of logical identity, *diversity*, and Hans Reichenbach's[5] *genidentity* (in one sense of that term): the property of self-identity through time. Now we can see how attempting to discuss individuality in terms of haecceities muddies the water. The problem is that they can be implicated in all its distinct aspects. One might argue that atoms are distinct because they possess distinct haecceities, or that instances are genidentical because they possess the same haecceity. Diversity and genidentity, however, could also be brute features of individuals. If we drop the issue of haecceities, we can treat the different aspects of individuality separately. This is as well because, as I shall show at the end, they are not equivalent. For now, let us concentrate on haecceitism.

## II. REPRESENTATIONS

Consider Democritean worlds, those worlds consisting of basic elements which we can call *atoms*: our qualitatively impoverished coin worlds or the point-particle world of classical mechanics for instance. In this section, I shall discuss what possible states such a world might have: or more precisely, what states theories of such a world might represent there to be. Since "possible states" are nothing but possible worlds, the structure in which states are represented by a theory is the structure of the universe of possible worlds postulated by the theory. The point of view of this paper is that haecceitism is thus a doctrine concerning such structures.

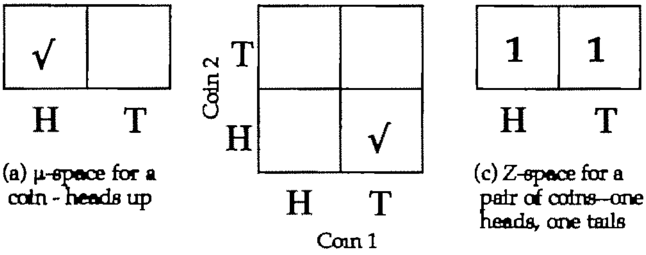
---

[4] Clearly, for such worlds to be qualitatively identical, haecceities must be excluded from the roll-call of qualities, perhaps on the grounds that they do not feature in explanations of mature science.

[5] *The Direction of Time* (Berkeley: California UP, 1956), pp. 270-71.

First we treat the possible properties of a single atom: let there be *m* logically independent properties, each with its own range of possible values—its spectrum. I shall take it that any spectrum can be represented on some portion of some number line. The obvious paradigm is spatial location represented by real coordinates.

The state of an atom is the conjunction of the properties that it possesses at an instant. This suggests a way of representing the states; we form the *m*-dimensional single-particle state space, *µ-space* (for gas 'molecule' in traditional statistical mechanics), as the Cartesian product of the number-line segments. Different states are represented by different points in µ-space, or, when more convenient, by their *m* coordinates in the space. For instance, the discrete spectrum heads-or-tails has two µ-states, which are represented by two cells (figure 1a). In classical mechanics in three-space, a particle[6] has six independent dynamic properties, which may be chosen to be its spatial coordinates and the components of its momentum, and a state is represented by a point in six-space, that is, by six coordinates. The dimensionality of µ-space is not important for present purposes; all we require is that for any atom, it contains a fixed number of distinct cells representing distinct states.



(a) µ-space for a coin - heads up

(b) Γ-space for a pair of coins—coin 1 tails, coin 2 heads

(c) Z-space for a pair of coins—one heads, one tails

Figure 1: State Spaces. More complex systems are represented in just the same way, except µ- and Z-spaces are larger, and Γ-space has more dimensions.

---

[6] Assuming that there are no internal degrees of freedom

Atoms may also posses constant properties, such as mass and the charge, but we shall be concerned only with homogeneous gases, composed of atoms of a single "species," all with the same constant properties and identical μ-spaces. We shall thus be able to ignore such properties: the atoms in the following discussion will differ qualitatively just in case they have different μ-states (in identical μ-spaces).

One-atom worlds are pretty dull. What are the states of a system of many atoms, a "gas"? The obvious answer is that they are the conjunctions of the states of each atom. They can be represented by assigning a μ-space to each atom, so that the global state space for a system of *n* atoms, Γ-*space* (for 'gas'), is the Cartesian product of *n* μ-spaces. For instance, for two coins Γ-space is composed of four cells, {heads,tails}×{heads,tails} (figure 1b). Or again, the state of three static particles in three-space will be represented by a point in a nine dimensional *configuration space*. In classical statistical physics, where particles are described by position and momentum coordinates, Γ-space is the *phase space*. Following this convention, we can generically call states in Γ-space *phases*.

But there is another possibility for the states of a many-atom gas: the state of the system is not the conjunction but the set of all the atomic states instantiated by the system. The representation of this alternative kind of state simply describes how many atoms are in each μ-state, and so the state space has the same structure as μ-space: this space is Z- (Zeta-) or *distribution space*. For discrete spectra, we associate each cell with a value equal to the number of atoms in the corresponding μ-state. For instance, a state one-coin-heads-and-one-tails is represented by associating the value 1 with both the heads cell and the tails cell (figure 1c). Hence Z-states are not represented by points of Z-space but distributions over it. Similarly for continuous spectra: the Z-space of a system of particles is a distribution of particle numbers over a single particle position-momentum space. Indeed, we shall talk of Z-states as *distributions*.

Z-space then just lists what atomic properties are instantiated in what numbers, whereas Γ-space also describes which individuals possess which properties. For instance, when Γ-space says "particles 1, 4, and 5 are over here, and 2 and 3 over there," Z-space simply says "there are three particles over here, and two over there." What is immediately apparent from this comparison is that there is a many-one mapping of Γ-space phases into Z-space distributions. For instance, the distribution is the same if now the phase is "1, 2, and 3 are over here, and 4 and 5 over there" instead.

How "many-one" is this mapping? Consider properties with discrete and finite spectra. We suppose that an atom has $c$ available states, so that μ-space is divided into $c$ distinct multidimensional cells. A many-particle distribution can be expressed as the tuple $<O_1,O_2,...O_c>$, where $O_i$ is the *occupation number* of the cell: the number of atoms in state $i$. So how many phases have this distribution? The result is a standard piece of combinatorics, which is worth recalling.[7]

Let $N_\Gamma(O_1,...O_c)$ be the number of phases corresponding to a distribution $<O_1,O_2,...O_c>$. Then, if $n$ is the number of atoms in the gas:

$$\begin{array}{llcl}
N_\Gamma(O_1,...O_c) = & \text{ways of picking } O_1 \text{ atoms} & \times & \text{ways of picking } O_2 \text{ atoms} \\
& \text{out of } n \text{ to be in state 1} & & \text{out of } (n\text{-}O_1) \text{ to be in state 2} \\
\times & \text{ways of picking } O_3 \text{ atoms} & ... \times & \text{ways of picking } O_c \text{ atoms} \\
& \text{out of } (n\text{-}O_1\text{-}O_2) \text{ to be in} & & \text{out of } O_c \text{ to be in state } c \\
& \text{state 3} & &
\end{array}$$

But these terms are simply binomial coefficients, so:

$$\begin{aligned}
N_\Gamma(O_1,...O_c) &= n!/O_1!(n\text{-}O_1)! \times (n\text{-}O_1)!/O_2!(n\text{-}O_1\text{-}O_2)! \; ... \times O_c!/O_c!(O_c\text{-}O_c)! \\
&= n!/(O_1!O_2!...O_c!) = n!/\prod_{i=1}^{c} O_i! \qquad \text{eqn 1}
\end{aligned}$$

The validity of this formula for a pair of coins is easily checked; for instance, for $O_T = O_H = 1$, we get $N_\Gamma = 2!/(1!1!) = 2$ heads-tails states. The corresponding quantity for Z-space is trivially unity, for Z-states just are distributions. If $N_Z(O_1,...O_c)$ is the number of Z-states corresponding to the distribution $<O_1,...O_c>$, then:

$$N_Z(O_1,..O_c) = 1 \qquad \text{eqn 2}$$

Clearly, important modal considerations are making themselves felt at this point, for the values of these $N$s are the numbers of possible global states compatible with a single qualitative description. Thus, we have nearly reached a central question: Does the qualitative character of an atomic world uniquely determine it, or is it yet incomplete in some way? Specifically, is Z-space adequate for describing all phenomena, or is the extra structure of Γ-space required? In the next section, I shall investigate why this is a question of haecceitism.

---

[7] See, for example, D. ter Haar, *Elements of Statistical Mechanics* (New York: Holt, Rinehart & Winston, 1964), p. 27. The key notion is that of a *binomial coefficient* $(B,A)$ = the number of distinct subsets of $A$ elements contained in a set of $B$ elements = $B!/\{A!(B\text{-}A)!\}$. Also recall that while in general $A! = 1 \cdot 2 \cdot 3 \cdot ... (A\text{-}2) \cdot (A\text{-}1) \cdot A$, we have the special identity that $0! = 1$.

First consider how empirical consequences might follow from these abstract considerations. The simplest way is by equipping these spaces with a probability function, a measure, and the simplest choice is a uniform distribution: indifference over the states. Then the fact that $N_\Gamma \neq N_Z$ will in general lead to differences in frequencies of states, something that one might hope to observe. So we might be able to tell which of Γ-space and Z-space actually capture possibility simply by seeing which of these statistics actually obtains. I shall return to this point in section IV.

## III. THE METAPHYSICS OF PHASE SPACE

Drawing metaphysical implications from observed phenomena is a game with only one obvious set of rules: first, we determine the different modal properties of the universes of possible worlds described by different representations, and then we investigate whether the data favor one. In this section, I tackle the modal issue; in the next, I shall look at the data.

First, we need to confirm that a Z-state gives a complete account of the qualitative character of a world: any two states with the same distributions are qualitatively identical.[8] I shall assume that single particle states are qualitatively complete by construction. That is, just as before we considered coins whose only properties are heads and tails (not mint year, denomination, and so on), now we take it that any possible qualitative differences at all between atoms are reflected in their μ-spaces. There are no further qualitative differences: the atoms are not, say, also different colors.

Now, worlds could have qualitative characters not determined by distributions of physical atoms: for instance, the dualist one in which a thirsty Descartes yearns for the cold beer that he believes is in the refrigerator. But even in this sort of example, there are no physical properties that supervene on Γ-space states, but not on distributions, at least in worlds similar to ours. If this modest claim is true, then we can hold any nonsupervenient (and hence nonphysical) qualities fixed and then ignore them as we consider how worlds vary through universes of possibility; we shall not overlook any qualitative differences "arising from" the atoms. Hence, any two states corresponding to the same distribution have the same qualitative character, insofar as it is determined by atomic structure at all.

---

[8] A couple of cautions should be made here, for by looking at the structure of state space, I am not immediately discussing full diachronic possible worlds, but rather instantaneous "miniworlds." This is not an issue at this point; for example, if Z-states give a complete description of the instantaneous qualitative character of a world, then a series of Z-states gives a complete qualitative description of a diachronic world. One might try to qualitatively distinguish identical Z-states in diachronic worlds by the histories that produced them, but I shall block this move by focusing on deterministic systems, with unique trajectories through any Z-state.

This seems correct. We know of no properties that depend not just on there being atoms of certain masses and charges at *x*, *y*, and *z* at certain times, but also that it is a particular atom, delineated nonqualitatively, that possesses the charge and mass and occupies point *x*. It is hard to see how one could offer a general a priori argument against such a property, but it is equally hard to imagine what would prevent us from viewing the dependence as being, not on a particular atom, but on some new unique physical property of that atom.

Haecceitism then does not hold in the Z-universe, if we assume that the Identity of Indiscernibles applies between states. By the qualitative completeness of Z-space, any two qualitatively identical states have the same distribution, and thus by assumption are strictly identical. But a single world cannot differ from itself in de re representations concerning individuals, and hence haecceitistic differences are not possible.

On the other hand, Γ-space is far more promising, for it contains distinct yet qualitatively identical states; distinct phases may correspond to a single distribution. Further, we have naturally been understanding such distinctions as arising purely because, for instance, one phase represents that atom 1 is in state *a*, and the other that it is in state *b*. This is a difference in de re representation, so such distinctions are haecceitistic differences in Lewis's sense. We shall thus take it that the doctrine of haecceitism holds in Γ-space, and that empirical evidence for Γ-space would be empirical evidence for haecceitism. Conversely, evidence for Z-space would be evidence against haecceitism.

Now that we have all the metaphysics in place, we need to work on the other half of the basic metaphysical argument: that the data force a choice between Z- and Γ-spaces.

## IV. THE EMPIRICAL BASIS

I now want to add some physics to the bare representations. I shall assume a Newtonian dynamics in which the complete qualitative description can be given in terms of the continuously valued positions and momenta. The evolution depends on these properties only, not on labels. Thus Newtonian theory can be represented in Z-space or in Γ-space. The point is that the form of classical atomic dynamics does not immediately select one of the state spaces. We need some more physics: time-averaged statistics.

The consequences of statistical physics are of two kinds, ***statistical*** and ***entropic***, an important distinction here. In classical mechanics, the former evidence speaks for haecceitism and the latter against it. Now, the exact frequency with which a given distribution occurs is not directly observable, but must be inferred from statistical properties. As an idealization, however, I shall talk as if we have direct access to the frequencies. Since I show that this perfect knowledge does not select one of the representations—in realistic situations they concur—I can conclude that neither can the less ideal knowledge obtained in practice. Entropy is different; if there were facts about the value of entropy of a system, then knowing the values could be of help. The thrust of the argument will be that there are no such facts.

So, how do the statistical arguments work? The idea is that since there are many phases for any distribution, the two spaces predict different long-run frequencies. Once these are calculated, all one need do is observe which actually occurs. For the computation, first assume indifference in both spaces, so that the probability of a distribution will simply be the number of ways of realizing it divided by the total number of distinct states. Let there be Z distinct distributions altogether[9] and a total of Γ distinct phases in the whole of phase space. Then the frequencies of a given distribution, $<O_1, O_2, \dots O_c>$, according to Z- and Γ-spaces respectively are, using equation two:

$$F_Z(O_1, \dots O_c) = N_Z(O_1, \dots O_c)/\text{all distributions} = 1/Z \qquad \text{eqn 3a}$$

$$F_\Gamma(O_1, \dots O_c) = N_\Gamma(O_1, \dots O_c)/\text{all phases} = N_\Gamma(O_1, \dots O_c)/\Gamma \qquad \text{eqn 3b}$$

If these disagree for physical systems, then they effectively give us a way of empirically counting the number of physically possible worlds, and settling the issue of haecceitism.

In this specific context, let me first say something about the assumption of indifference over the possible states. One might point out that without the assumption of indifference, $F_\Gamma$ and $F_Z$ could be chosen to be anything, including identical functions. In this case, the argument cannot get off the ground, and there is no more evidence for haecceitism than for antihaecceitism. But to adopt this position is to step into a messy debate about prior probabilities. In the next section, I shall claim that the argument for haecceitism can be defeated much more solidly, so I shall not entertain this possibility.

I shall take it that the data of classical statistical mechanics as inferred from thermodynamics reveal that the correct frequencies are those given by $F_\Gamma$: equivalently, the statistics are Maxwell-Boltzmann. The *statistical argument* infers back from the observed statistics, arguing that they provide confirmation of the correctness of Γ-space, and hence of the veracity of the doctrine of haecceitism.

---

[9] For $n$ particles each with $c$ possible states, it is not hard to show that $Z = (n+c-1)!/n!(c-1)!$—see, for example, Reichenbach, p. 230.

I claim that this is the standard position, adopted, for instance, by texts on statistical physics (for example, ter Haar, pp. 71-72). After all, the Gibbs paradox,[10] which has exercised physicists so much, relies, in one of its forms, on the individuality of atoms. This position is also present in the views of certain philosophers: for instance, Michael Redhead and Paul Teller[11] make it part of a traditional conception of classical particles (*ibid.*, pp. 202-03), which they argue is inapplicable in quantum mechanics. Reichenbach suggests the same idea:

> ...in classical physics, two arrangements are regarded as different when they result from one another by an interchange of molecules between two given compartments.

But, in quantum statistics:

> It was shown by Bose and Einstein that...we must give up regarding two arrangements as different when the particles in different compartments are interchanged (*op. cit.*, pp. 228-30).

I shall say more about entropy later, but the important point is that it is a function of the number of distinct states, $N$ (= $N_\Gamma$ or $N_Z$), corresponding to a distribution: the entropy, $S = S(N)$. Since $N$ takes on different values in Z- and Γ-spaces, so will the entropy. Let us take it that there is prima facie evidence (to be discussed) from thermodynamics that $N_Z$ gives the correct value for the entropy: that is, $S = S(N_Z)$. The *entropic argument* infers back from this evidence, arguing that it provides confirmation of the correctness of Z-space, and hence of the falsity of the doctrine of haecceitism.

Now that I have clearly stated the two contrary arguments, statistical and entropic, I can proceed to see which, if either, prevails. Does the empirical evidence decide the metaphysical issue?

## V THE STATISTICAL ARGUMENT

The statistical argument is not deductively valid; all that is shown is that Γ-space (and indifference) yields the observed statistics. It does not immediately follow from the confirmation of Γ-space that Z-space is disconfirmed, unless one also shows that they disagree in the relevant experimental context. For the coins, they obviously do disagree (we get heads-tails either one time in two, or one time in three); but it is easy to see that in physically realistic cases the predicted frequencies agree. Hence the observed statistics confirm Z- and Γ-spaces, antihaecceitism and haecceitism, equally.

---

[10] For example, N. G. van Kampen, "The Gibbs Paradox," in W. E. Parry, ed., *Theoretical Physics* (Oxford: Pergamon, 1984), pp. 303-12.

[11] "Particle Labels and The Theory of Indistinguishable Particles in Quantum Mechanics," *British Journal for the Philosophy of Science*, XLIII (1992): 201-18.

The coins are irrelevant to settling the issue because, of course, one needs to consider atoms whose μ-spaces are qualitatively complete, unlike heads-tails for real coins. The kind of real system whose atoms have qualitatively complete μ-states (bar shared constant properties) is a homogeneous classical gas of rigid Newtonian particles. The key feature of such a system which ensures the agreement of phase and distribution is—as we shall see below—that no two Newtonian particles ever simultaneously occupy the same point of μ-space: $O_i = 1$ or 0 always.

Result I: Given $O_i \leq 1$, distribution and phase spaces entail the same frequencies.

Proof:

(i) Assume $O_i \leq 1$ for every single particle state, so $O_i! = 1$ for every $i$. Then, using equation one:

$$N_\Gamma(O_1,...O_c) = n!/\prod_{i=1}^{c} O_i! = n!/(1 \cdot 1...1) = n!$$

Given the assumption, there are exactly $n!$ phases to any distribution, and hence if there are Z distributions, the total number of phases, Γ, is $n!Z$.

(ii) Hence, by (i) and equations three:

$$F_\Gamma(O_1,...O_c) = n!/n!Z = 1/Z = F_Z(O_1,...O_c)$$

Given the assumption, the statistics implied by both Z- and Γ-spaces for distributions are identical.[12] In this situation, statistics speak equally for both spaces, and do not force any choice. This is significant, for there are two good reasons to suppose the assumption is correct.

First, it holds if atoms are impenetrable, for then they just cannot occupy the same location. But less problematically, it is also true if one of the atomic properties is characterized by a number line with an infinite number of points, as long as there are a finite number of atoms (and given the assumption of a—Lebesgue—uniform probability). The classical world is spatial, so all position and momentum coordinates are of this kind. We can see the relevance of this very simply: consider for illustration a system of two particles in one-space dimension. The measure of the set of cells of double occupancy is the measure of the "diagonal," $x_1 = x_2 \cap p_1 = p_2$, in the four-dimensional phase space. In the limit that the cells become continuous points, this line has measure zero in phase space, and similarly for higher dimensions. The point then is not that the spaces necessarily are so constrained, but that they might as well be for the statistical properties; which is exactly what we need to see that empirical results will not distinguish the spaces for us. Of course, other sets of points have measure zero, too, but the point here is that this set is the only one in which Z- and $\Gamma$-space counting disagrees by anything other an irrelevant ($n!$) counting constant.[13]

---

[12] Note that one need not assume indifference to derive these results, merely that the prior probability, $p(O_1,...O_c)$, is the same function of distributions in both spaces. Note, too, that exactly the same results apply for functions of distributions: given the assumption, $f(O_1,...O_c)$ will take a given value just as often in Z-space as in Γ-space. For a proof, see my *The Philosophy of Fields and Particles in Classical Mechanics, including the Problem of Renormalization* (Ph.D. Thesis, Rutgers University, 1995).

Those familiar with statistical physics may find this conclusion unsettling. When one deals with thermostatistics one works in phase space, and seems to make no assumption that occupation numbers are either one or zero. Proceeding in this way, it seems obvious that Z-space is just wrong (as it is for the coins). But now we have seen that, as a mathematical consequence of a true observation, our opinion here is incorrect and Z-space is equally appropriate. There is a tension produced by the gap between proving that the approaches agree and explicitly seeing how. Elsewhere,[14] I derive the standard formulae of classical statistical mechanics from Z-space, showing that there is indeed no disagreement. Here a simple result will make the same point.

First note that, in general, thermostatistics is based, not on counting states characterized over a space of individual cells, but over a *coarse-grained* space in which cells are grouped into blocks. So suppose that our μ-space contains $b$ blocks each containing $d$ cells, for a total of $b \cdot d = c$ cells. Let $<n_1, \ldots n_b> \equiv <n_i>$ be the distribution over the blocks, as the $O_i$s are over the cells. Clearly, even if $O_i \leq 1$, it does not follow that $n_i \leq 1$: our constraint on cell occupation numbers is compatible with the absence of constraints in the familiar coarse-grained approach to thermostatistics.

---

[13] In quantum mechanics, spectra get quantized, so that they are no longer continuous, and the assumptions of the proof fail: states with $O_i > 1$ are no longer improbable. In this case, the possibility of settling this metaphysical issue experimentally reappears. For a full investigation of this suggestion, see my "Identity Quantum Mechanics and Common Sense," *The Monist*, LXXX, 1 (1997): 118-30.

[14] *The Philosophy of Fields and Particles in Classical Mechanics*, pp 74ff.

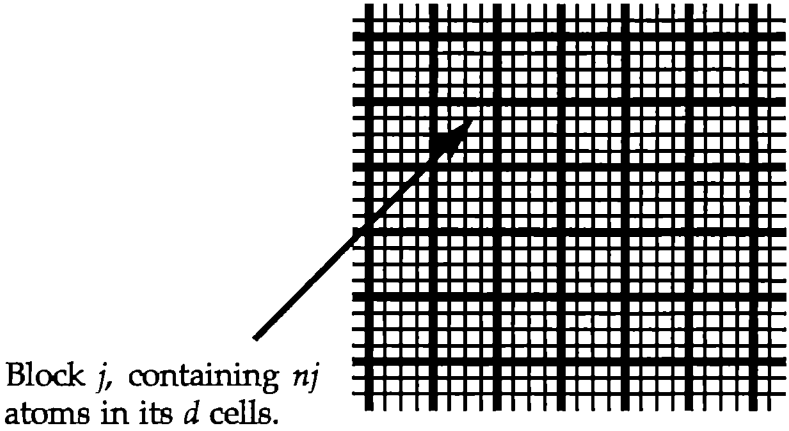


Figure 2. A region of Z-space coarse-grained into equal blocks.

Strictly speaking, the empirical evidence of thermodynamics to which we have referred is that $N_\Gamma(n_i)$ ($= n!/\prod_{i=1}^{k} n_i!$ from equation one) gives the correct—'Maxwell-Boltzmann'—number of states corresponding to a given *coarse-grained distribution* over the blocks. I shall now calculate $N_Z(n_i)$, the number of fine-grained distributions corresponding to a coarse-grained one, and show that it agrees with the empirical $N_\Gamma(n_i)$, up to an irrelevant constant that vanishes in the observable frequencies. That is, when we take the standard blocked approach to statistical physics, the same result is obtained whether we suppose the underlying fine-grained space to be phase or distribution space. (Note that $N_Z(n_i) \neq 1$ because many distribution over cells will give the same coarse-grained distribution.)

In this derivation, we again assume $O_i \leq 1$, which we take to follow from a sufficient diffuseness that $d \gg \sum_i n_i = n$, so that each block has many more cells than the total number of atoms in the gas. Again, this is justified on physical grounds by an appeal to the vast number of positions and momenta available to a particle in any finite region. Then:

Result II: Given $O_i \leq 1$ and $d \gg n$, $F_Z(n_i) \approx F_\Gamma(n_i)$

Proof:

(i) By $O_i \leq 1$, each fine-grained cell will contain at most one atom, so[15]

---

[15] This proof can be given without an explicit assumption that $O_i \leq 1$, using only diffuseness, $d \gg n$. Then, to the initial expression for $N_Z(n_i)$ add terms, for example, for picking $n_\Gamma$-1 cells out of *d* (that is, for double occupancy), for $n_\Gamma$-2 cells out of *d*, and so on. But the binomial coefficients for such terms are at least of order $d!/\{(n_i$-

$$N_Z(n_i) = \underset{\text{cells out of } d}{\text{ways of picking } n_1} \times \underset{\text{cells out of } d}{\text{ways of picking } n_2} \ldots \times \underset{\text{cells out of } d}{\text{ways of picking } n_b}$$

$$= \prod_{i=1}^{b} d!/\{n_i! \cdot (d-n_i)!\}, \text{ using binomial coefficients (see footnote 7).}$$

But for $p \gg q$, $p!/(p-q)! \approx p^q$, so using $d \gg n$ (diffuseness) we find, by equation one:

$$N_Z(n_i) \approx \prod_{i=1}^{b} (d^{n_i}/n_i!) = d^{n_1+n_2+\ \ n_b}/\prod_{i=1}^{b} n_i! = d^n/\prod_{i=1}^{b} n_i! = (d^n/n!) \cdot N_\Gamma(n_i)$$

(ii) We can write the total number of (fine grained) distributions as $\sum_{\langle n_i \rangle} N_Z(n_i)$ to show that it is the sum of $N_Z(n_i)$ over all coarse-grained distributions. Similarly, $\sum_{\langle n_i \rangle} N_\Gamma(n_i)$ is the total number of fine-grained phases. Hence, using equations three and (1):

$$F_Z(n_i) = N_Z(n_i)/\sum_{\langle n_i \rangle} N_Z(n_i)$$
$$\approx \{d^n/n! \cdot N_\Gamma(n_i)\}/\{d^n/n! \cdot \sum_{\langle n_i \rangle} N_\Gamma(n_i)\} = N_\Gamma(n_i)/\sum_{\langle n_i \rangle} N_\Gamma(n_i)$$
$$= F_\Gamma(n_i)$$

The point of the last step is that constant differences in counting wash out in frequencies, which are all that can be observed, and an (albeit large) constant is all that separates $N_Z$ from $N_\Gamma$. Here it is clear that the $1/\prod_i n_i!$ factor characteristic of Γ-space counting and Maxwell-Boltzmann statistics also appears in Z-space statistics, not because we count over permutations of μ-states among atoms (as in the derivation of equation one), but because it is introduced by the binomial coefficients for arrangements within blocks. Thus there is no mystery at all why, in this case, the two spaces are in agreement. Note finally, that although the expression is approximate, in the case of a classical gas, we can take the system to be arbitrarily diffuse, and so there will be arbitrarily high experimental agreement.

The lesson should be crystal clear. The choice of space is not empirical, because once $O_i \leq 1$ is imposed, the remaining Γ-states can be arranged into $n!$ sets which map 1-1 into Z-states: $n!$ is just an irrelevant normalization constant which washes out in the physically relevant frequencies. The philosophical pay-off is that statistics do not select one space over the other, and hence will not yield conclusions

---

$1)!(d-n_1+1)!\} = \{n_1/(d-n_1+1)\} \cdot d!/\{n_1!(d-n_1)!\}$, which by diffuseness is highly suppressed by a factor $\approx n_1/d$. Such terms can be discarded and the expression for $N_Z$ is an approximation, not an identity.

about modal metaphysics. Rather surprisingly, the Maxwell-Boltzmann distribution does not entail a haecceitistic universe. The statistical argument fails. What about the entropic argument?

## VI. THE ENTROPIC ARGUMENT

The entropy in statistical mechanics is defined by the logarithm of the number of available (coarse-grained) states:

$$S = k \cdot lnN(n_i) \qquad \text{eqn 5}$$

where $k$ is Boltzmann's constant. Now, statistical physics has empirical import because it is intended to explain thermodynamics. In this reduction, the statistical entropy, $S$, is identified (at least in equilibrium) with the thermodynamical entropy $\sigma$. The main evidence in favor of $S_Z$ (obtained when, in equation five, $N = N_Z(n_i)$) is that $\sigma$ is supposed to be *extensive* or *additive*. That is, if two equilibrium systems are in mutual equilibrium, then the entropy of their union is the sum of the entropy of the two parts: $\sigma_{1\cap 2} = \sigma_1 + \sigma_2$. Via the identification we expect for the statistically defined equilibrium entropies that $S_{1\cap 2} = S_1 + S_2$. One can straightforwardly check that this equality holds for $S_Z$, but not for $S_\Gamma$, obtained by substituting $N_\Gamma(n_i)$ for $N$ in equation five. The nonadditivity of $S_\Gamma$ thus is supposed to rule out Γ-space in favor of the extensive $S_Z$. Further, observation of entropy-dependent properties shows that they are in agreement with $S_Z$. Hence we are supposed to accept Z-space and antihaecceitism. I oppose this conclusion in two ways: on the one hand, there is nothing compelling about the inference from $S_Z$ to Z-space; on the other, the data of thermodynamics do not speak unequivocally for $S_Z$.

Let us assume that $S_Z$ is the correct expression of the statistical entropy; in the sequel, I shall show this assumption to be unwarranted. Entropy is a function of the distribution, but any Γ-state uniquely defines a distribution so the entropy can equally be considered a function with either Γ-space or Z-space states as its domain. Assuming that $S_Z$ is the correct expression of entropy, then the two functions are easy to find from result II[16]: in Z-space $S_Z = k \cdot lnN_Z$, and in Γ-space $S_Z = k \cdot ln(cN_\Gamma/n!)$. In this seemingly trivial sense either space supports any definition of entropy. The question is: Is there some significant sense in which Γ-space does not support $S_Z$?

The only apparent response is to give a physical meaning to the entropy that requires the assumption of Z-space. For instance, the statistical entropy can be implicated in an explanation of the "inevitability" of equilibrium. By equation five, $S_Z$ increases with $N_Z(n_i)$. The entropy grows as the system moves into distributions corresponding to more numerous sets of states. Since (given indifference) these states are more likely, it follows that entropy will "most likely" increase to its maximum.[17] Thus $S_Z$ has a physical meaning as a measure of the likelihood of a distribution, but this meaning arises interchangeably from Γ- or Z- space. It follows from our results that the likelihood of any distribution is the same in Γ- or Z-spaces, and hence that they can assign the same entropy to any distribution and preserve the physical meaning.

---

[16] Note carefully that I am amending the definition of $S$. To make the same correction Gibbs inserted an $n!$ factor in $N_\Gamma$, and concluded that he had an argument for Z-space. My point is that one is free instead to redefine the relationship between $S$ and $N$, so that no such conclusion is warranted.

The point here is that defining $S_Z = k \cdot ln(cN_\Gamma/n!)$ is no mere sleight-of-hand, for the physical meaning of entropy does not speak in favor of Z-space. There is nothing in this story of statistical mechanics to force a commitment to one space or another, or to any stance on haecceitism. But not only is there no valid inference from $S_Z$ to anti-haecceitism, there is in fact no reason for selecting $S_Z$ over $S_\Gamma$.

First consider an argument by N. G. van Kampen (*op. cit.*) to the effect that additivity is not necessary for thermodynamics. Assume that the laws of thermodynamics entail all the relevant observable phenomena. The question then is: Is additivity entailed by the laws? If not, then it is not a consequence of conventional empirical thermodynamics. I restrict our attention further by observing that the thermodynamical entropy is introduced through the Second Law, $d\sigma = dq/T$ (where $dq$ is the increase in heat and $T$ is the temperature). The important point about the law for the argument is that it only defines entropy differences as work is done, not an absolute value.

First, does the Second Law entail the additivity of the entropies of distinct disjoint systems? Since it only describes a single system at different points in an evolution, the Second Law says nothing at all about extensivity in this case. What about additivity in the situation where two samples are physically joined? Surely, the evolution from two physically disjoint samples into one can be considered as the development of one system. But this does not circumvent the problem, for, according to van Kampen (*op. cit.*, p. 305), deriving additivity in this way relies on the assumption that the entropy of identically prepared distinct systems is equal. Again, although this is a reasonable and helpful definition, it cannot follow from a law concerned solely with differences along thermodynamical trajectories. The answer seems to be that additivity is not an observable consequence of thermodynamics.

---

[17] The sense of 'most likely' is subtle: see, for example, Lawrence Sklar, *Philosophy of Physics* (Boulder: Westview, 1992), pp. 113-17.

KP-836

I conclude that there are no observable entropic properties sensitive to the difference between Γ- and Z-space, and hence no observation will give us any decisive insight into haecceitism.

van Kampen (*op. cit.*, p. 307) explains why we should not be surprised about this negative conclusion. The idea of statistical mechanics is that the data be completely explained in terms of the statistics of microstates—after all, if the world is made up of atoms, it is what they do that counts. If this principle is correct, then predictions derived using the entropy cannot go beyond the statistical consequences of the theory. But, as we have clearly seen, all the statistical consequences are compatible with both haecceitism and antihaecceitism; the observable entropic consequences must be too.

## VII. CLASSICAL CONCLUSIONS

So far I have established the negative result that the data of thermodynamics do not decide the metaphysical issue of haecceitism; now I should consider the possibility that we are faced here with an issue that is "metaphysical" in the traditional sense of transcending mere scientific considerations. The point is that the difference between Γ- and Z-spaces is genuinely physically redundant. Certainly, they are empirically underdetermined, but it is not simply as if one of the theories had extra unobservable processes taking place, rather phase-space simply asserts a difference where distribution-space sees none. That is, Γ-space asserts exactly everything that Z-space does, plus it claims that every kind of haecceitistic difference is possible, and nothing more. Given this, one is faced with the question of why such surplus representational power should be postulated. At the very least it provides a significant sense in which the choice between Γ- and Z-spaces is not between physical theories, but between metaphysical outlooks. And arguably, since the extra claim of haecceitism does no work, Okhamite considerations should drive us to reject it.

I am tempted to go further here and claim that this is no real difference at all: that this so-called metaphysical difference is nothing but empty words. It simply asserts a difference without giving an account of what it is supposed to be. To say that "this atom" now has this property, and now that, and for this to be saying something substantial presupposes that something meaningful is being done in the denotation in the different situations, and the burden of proof falls to the haecceitist to show that this is so.

Finally, let me put the example of classical statistical mechanics to work to discover the logical relations between the various aspects of individuality. In the first place, the assertion of genidentity and diversity is compatible with the denial of haecceitism. An "excluding" property like location in μ-space clearly behaves in some ways akin to a haecceity, grounding a kind of qualitative genidentity and distinctness (though not haecceitism.) And there is no prima facie reason why one should not suppose that these qualitative features of individuals co-occur with metaphysical ones. Atoms distinguished by locations can be supposed metaphysically distinct individuals, and instances of particles joined by continuous space-time trajectories instances of a strictly identical individual. (Of course, these features are also compatible with the rejection of any extra metaphysical facts, so we are not forced into any metaphysics here either.) This reasoning holds in both Z- and Γ-space, and so antihaecceitism does not rule out genidentity or diversity. Neither diversity nor genidentity are sufficient for haecceitism.[18]

But metaphysical genidentity and diversity do not necessarily hold in virtue of qualitative features; as an alternative, distinct trajectories can be thought of merely as the "marks" of these metaphysical features. But in metaphysics, surely something can occur without its qualitative marks? In this case, in the face of the denial of haecceitism, these components of individuality do not depend on the existence of suitable properties. But this is to say that individuals can be distinct and genidentical regardless of what properties they have in an antihaecceitistic universe.

What about the reverse implications? Haecceitism must imply diversity: How can we even talk about what is represented de re about Peter unless he is distinct from Paul? Genidentity is another matter, for surely worlds in a haecceitistic universe could be composed of individuals constantly created and destroyed every instant. Similarly, genidentity implies diversity, for how can an atom be reidentified unless it can be picked out of an ensemble?

An important motivation for carrying out this lengthy investigation is to help us pinpoint the ways in which quantum mechanics differs from classical mechanics. One could easily be led to think that an important difference follows from the different ways in which "identical" particles are handled in the two approaches. Different statistics point to different metaphysics of individuality in each case: haecceitism in classical mechanics and antihaecceitism in quantum mechanics. My analysis has been directed at showing that there are no very heavy metaphysical implications of classical physics, and that therefore anticipated innovations in the notion of an individual in quantum mechanics will not be innovative at all. This is not to say there are no differences at all, but it is to say that the novelty of quantum mechanics is artificially heightened if we add to classical physics the unwarranted baggage of "a metaphysical position which has been moribund for centuries."[19]

---

[18] Actually the issue is not quite this straightforward if paths of genidentity can become detached from a continuous trajectory, so that continuents 'hop around' the continuous paths mapped out by matter. Then any instantaneous state is haecceitistic, because 'this atom' could have started on any of the trajectories.

NICK HUGGETT

University of Illinois/Chicago

---

[19] Bas van Fraassen, *Quantum Mechanics: An Empiricist View* (New York: Oxford, 1991), p. 434. For an investigation of the metaphysics of identity entailed by quantum mechanics, see my "Identity: Quantum Mechanics and Common Sense."

# SEEING INTERSECTING ECLIPSES*

I am viewing a double eclipse of the sun. Traveling east is the heavenly body Far. Traveling west and nearer to me is the smaller body Near. Near is close enough to compensate exactly for its smaller size with respect to shadow formation. Near and Far look the same size from my vantage point. When Near falls exactly under the shadow of Far, it is as if one of these heavenly bodies had disappeared. Do I see Near or Far?

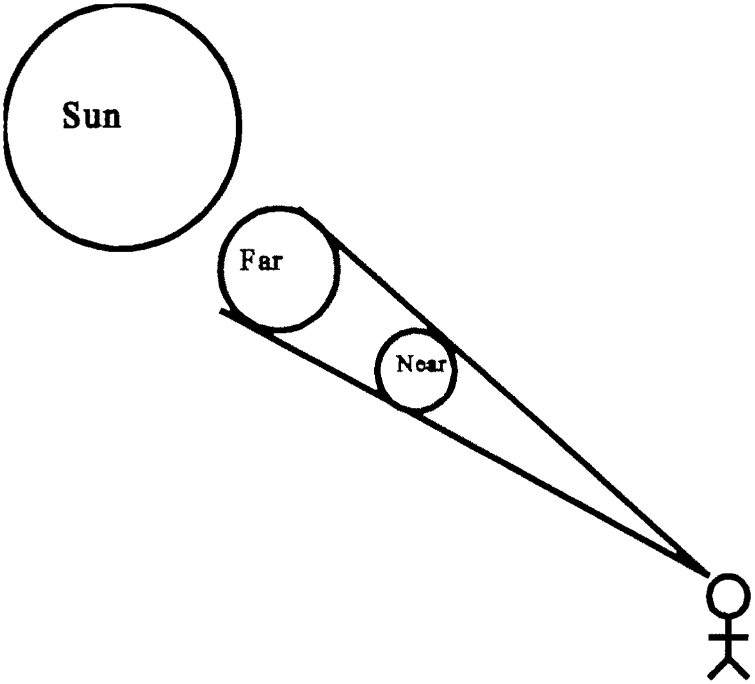


Figure 1

---

* I thank far-sighted members of colloquium audiences at Dartmouth College, the University of Connecticut, and the University of Delaware for their perceptive comments. I also received insightful counsel from Jonathan Adler, John Carroll, Fred Dretske, Julia Driver, Doug Ehring, Brian McLaughlin, David Robb, Walter Sinnott-Armstrong, Gerald Vision, and Denis Walsh.

0022-362X/99/9601/25–49

Common sense answers that I see Near rather than Far: Near is an opaque body that completely blocks my view of Far. Since I see something, I see Near.

## I THE CAUSAL THEORY CORRECTLY PRECLUDES NEAR

According to the causal theory of perception, *S* sees object *O* just when there is an appropriate causal connection between *S* and *O*. 'Appropriate' has proven difficult to spell out. But the theory entails a necessary condition that clearly precludes the possibility that I see Near: *S* sees *O* only if *O* is a cause of what I see. Near is totally within the shadow cast by Far. An object that is completely enveloped in a perfectly dark shadow cannot be seen. A cat silhouetted against the moon can be seen by virtue of the light it blocks. But since Far is preventing any light from striking Near, Near is not casting a shadow. Near is causally idle and therefore invisible. If light were being reflected by Near into my eyes, Near would preempt Far; and I would see Near. But since I am actually seeing by virtue of partial light blockage, it is Far that preempts Near.

H. P. Grice[1] motivated the causal theory of perception with a thought experiment. Suppose I believe I am facing an unoccluded pillar. Unbeknownst to me, there is a mirror that is reflecting the image of the pillar to my rear. On the opposite side of the mirror there is a pillar, indeed one that would look just like the one in the mirror. If the mirror were removed, there would be no phenomenological difference. Under these conditions, common sense correctly says I see the mirror and not the occluded pillar. The reason is that the occluded pillar plays no causal role in what I see.

Or suppose that there are two identical coins in my line of sight, one behind the other. If the front coin is rolled away, the scene will look the same. Nevertheless, I do not see the rear coin because it makes no actual difference to what I see. The reason why I see only the objects facing me is that only they are in a position to transmit light into my eyes. More generally, the perceiver's field of vision is widely agreed to comprise the *front* surfaces of those opaque bodies lying within a conical region radiating outward from the perceiver's open eyes. The causal theory of perception provides a natural explanation of why my field of vision is structured as it is.

Alvin Goldman[2] uses satisfaction of this environmental relation between perceiver and perceived as one of three defeasible factors in determining whether an object is seen. His second factor is the existence of a physical mechanism. By this Goldman means a causal pathway by which the object transfers energy to the perceiver. His last factor is counterfactual dependency: Would the scene have looked different had the object not been there? Each of these factors has some say but none has a veto. Consequently, there will be hard cases. Goldman considers Jaegwon Kim's example of an astronomer looking at a black hole through a telescope. The astronomer sees the black hole by virtue of its contrast with the background star light. According to Goldman, the black hole is in the perceiver's field of vision and satisfies the counterfactual dependency condition but violates the physical mechanism condition.

---

[1] "The Causal Theory of Perception," *Proceedings of the Aristotelian Society*, Suppl. Vol. xxxv (1961): 121-68, here p 141.

[2] "Perceptual Objects," *Synthese*, xxxv (1977): 257-84.

To save the causal theory, Goldman claims that 'cause' is ambiguous between a counterfactual-dependency sense and a physical-mechanism sense. Since parallelism and occasionalism deliver counterfactual dependency without causation, this ambiguity defense fails. For good or ill, the causal theory unequivocally condemns idle objects to invisibility. If God made an astronomer have hallucinations of a black hole that reliably track the black hole, the astronomer would not see the black hole.

Since black holes violate just one of Goldman's conditions, he (correctly) acquiesces to the intuition that the astronomer sees the black hole. It is striking that Far violates all three of Goldman's conditions. The environmental relation is unsatisfied because Near is in front of Far. The counterfactual dependency condition is violated because I would notice no change in the scene if Far were not present. And the physical mechanism condition is violated because Goldman requires that the physical mechanism involve an energy transfer from the object to its perceiver.

Although Near satisfies the environmental condition, Near violates Goldman's other two conditions. Near transfers no energy to me. And if Near were absent, the scene would be indistinguishable. Hence, Goldman is committed to saying that I see neither Near nor Far.

I have future comments on each of Goldman's factors of perception. But here is an immediate remark about the environmental relation. The relativity of 'occlude' forces a corresponding relativization of our field of vision. Consider two dark nebulae that exactly overlap in the manner of Near and Far. A dark nebula is a dense cloud of interstellar dust. Dark nebulae do not transmit light in the optical region of the spectrum. Yet they are conspicuous in any photograph of the Milky Way because of their contrast with ambient starlight. The twist is that "dark" nebulae glow brightly in the infrared because interstellar dust efficiently absorbs optical and ultraviolet radiation. An astronomer peering through an infrared telescope can see the near nebula. But an astronomer using an optical telescope cannot. Relative to infrared observation, the near nebula occludes the far nebula. But the reverse holds for optical observation. The infrared astronomer has a classical field of vision in which he sees the fronts of objects. The optical astronomer has a back-lit field of vision in which one of the objects is visible by its backside.

The correct solution to the eclipse riddle is that I see Far. This answer follows as a bold consequence of the causal theory of perception, given natural background assumptions. 'I see Far' will seem bizarre as long as we persist in modeling all vision on seeing objects that transmit light (by reflection, refraction, or emission). We must be careful not to prejudice the explanation of our field of vision by overlooking alternate lighting conditions. In particular, when objects are back-lit and are seen by virtue of their silhouettes, the principles of occlusion are reversed. In back-lit conditions, I can hide a small suitcase by placing a large suitcase *behind* it. The reversal runs deeper: we see the backs of back-lit objects. Although this thesis seems like parapsychology, my premises are drawn from the reigning orthodoxy. I am a conservative rather than a radical.

## II. SHADOW ARCHITECTURE

Any opaque object can be seen by virtue of its silhouette. The shade formed by a back-lit object is a volume that commences from the far side of the object. Call this initiating surface the silhouette.[3] The shade extends from the silhouette in accordance with principles of projective geometry. The shadow can be truncated only by light. Intervening opaque objects do not stop the shadow. After all, the shadow is a privation of light rather than a positive force. Hence, it does not need to penetrate objects to reach the opposite side. Nevertheless, the shadow is sensitive to whether the object contains light. Consider a silhouetted safe that contains a flare. If the flare is burning, then the *near* wall of the box blocks light and so casts the shadow at the base of the safe.

The dark area on the initial surface of an intermediate object is called a "cast shadow" by architects. They call the shade between the beginning and end of the shadow "the invisible shadow" (though they admit it can be seen when the air is dusty).

---

[3] The representational sense of 'silhouette' is etymologically prior to its mereological sense. The notorious French controller general of finance, Etienne de Silhouette (1709-1767), had a hobby of cutting shadow portraits out of black paper. 'Silhouette' came to be applied to depictions of this sort and then later became applied to the causal basis of those depictions. Compare 'silhouette' with 'pin-up', which at first denoted only poster depictions of sexy women and then came to denote the subjects of those depictions.

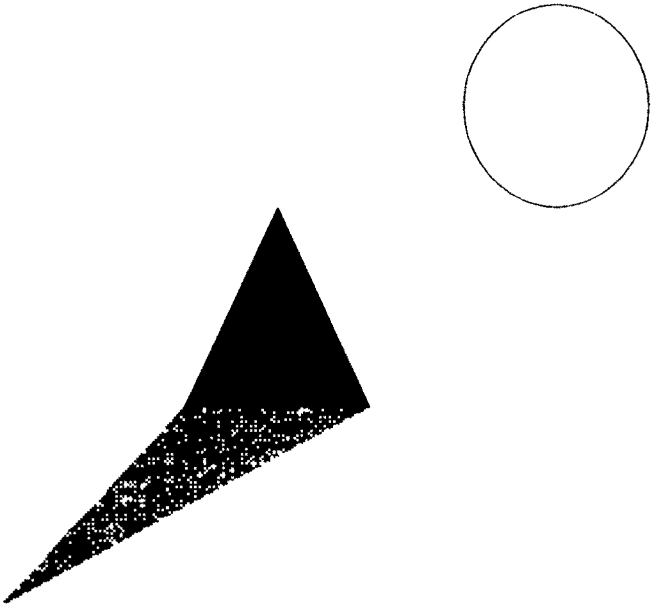


Figure 2

Shadows are not mere absences of light. Shadows require light sources and transparent regions where light would have entered were it not for objects that block the source. A brick resting flat on the ground at high noon casts no shadow. If I lift the brick, it casts a shadow. But I have not discovered a pre-existing shadow.

Philosophers exacerbate the mystery of shadows by treating the shadow as if it consisted solely of the cast shadow. For instance, Avrum Stroll[4] contrasts the three dimensionality of a physical object with the two dimensionality of shadows. Whereas two objects cannot be the same place at the same time, a two-dimensional shadow can be "located in the same portion of space that, say, a stretch of road is occupying. So shadows have an existence in space without displacing space" (*ibid.*, p. 75). Stroll is led to wonder whether shadows violate John Locke's[5] principle that two things of the same kind cannot be in the same place at the same time. If two people cast shadows that exactly overlap, why must there be one shadow rather than two (*ibid.*, p. 72)?

---

[4] *Moore and Wittgenstein on Certainty* (New York: Oxford, 1994).

[5] Locke states the principle in *Essay concerning Human Understanding*, Essay II, chapter 27, section 1.

Puzzlement is diminished by the architect's conception of the cast shadow as a surface of a three-dimensional volume. The two dimensionality of the cast shadow is due to its status as a surface rather than its status as a shadowy thing. Hence the cast shadow has no *special* property of being in space without displacing space. If an interloper walks exactly into my invisible shadow, then he does not cast a shadow. To cast a shadow, he must block light.

With the help of separate light sources, two coins can causally overdetermine the same cast shadow. If you glue penny *A* to the center of flashlight *A* and penny *B* to the center of flashlight *B*, then you can position the pair of flashlights so that penny *A* and penny *B* cast the same shadow against a wall. The symmetry of the situation prevents either flashlight from preempting the other. Since each penny generates a distinct volume of shade, their shared cast shadow is like the shared section of two intersecting roads. It is also possible for the two pennies to share nearly all of their invisible shadows. Just point the two flashlights directly at each other.

The parasitic nature of shadows makes it natural to infer that "a shadow is sustained in existence by the continued existence of its originating source."[6] But shadows can survive the destruction of their originators; consider a tree that is constantly illuminated as it petrifies into stone. The shadow begun by the tree is continued by the stone.

It is also possible for a shadow to persist without any sustaining object. Light travels at 299,792,458 meters per second in a vacuum. The moon is about 384,400,000 meters away from the earth. Hence, if the moon were instantly obliterated during a solar eclipse, its shadow would linger for over a second on the surface of the earth. If the moon were further away, its shadow could last several minutes. We can extrapolate to posthumous shadows that postdate their objects by millions of years. We can also speculate about an infinite past in which a shadow is sustained by a beginningless sequence of objects. As one object is destroyed, it is seamlessly replaced by an object of the same shape and size. Arguably, this shadow antedates any object in the sequence and so threatens the principle that every shadow is caused by an object.[7]

---

[6] Samuel Todes, "Shadows in Knowledge," in D. Ihde and R. M. Zaner, eds., *Selected Studies in Phenomenology and Existential Philosophy* (The Hague: Nijhoff, 1975), pp. 95-113, here p. 96.

[7] Yet more speculative is Samuel Levey's scenario involving an infinite sequence of walls where wall *n* has a thickness $1/2^n$ meters. If the sun rises from the thin side, which wall casts the shadow at the base of the thickest wall? I suspect Levey's scenario is either physically impossible or underspecified. What do the photons hit if there is no first wall to strike?

Do we see silhouetted objects that no longer exist? The answer is the same as the solution to the old riddle prompted by starlight that survives its star: Can we see stars that no longer exist? Yes, but we see them as they were, not as they are.

Since the silhouette of an object is part of the object, there can be cast shadows without silhouettes. Thus the leading edge of a shadow need not be a silhouette. (Since an uninterrupted shadow ends at a region of space instead of at the surface, it follows that some shadows have neither a silhouette nor a cast shadow.) When the object is destroyed, the leading edge of the dark three-dimensional volume races toward its opposite edge at the speed of light. Small volumes of shade appear to be destroyed at exactly the time that the object is destroyed. But, strictly speaking, shadows last longer than their destroyed objects.

The speed of light compels a surprising qualification to my answer to the eclipse riddle.[8] Initially, Near blocks light because there will be photons trapped between Far and Near. Once Near absorbs these intermediate photons, it will no longer block any light and consequently become invisible. Only at that stage do we see Far.

Although cast shadows are displaced and usually distorted, silhouettes provide a visual match that is comparable to that achieved with transmitted light. Seeing a silhouette of an object counts as seeing the object itself. This principle is prehistoric and culturally universal. Silhouette portraits are the earliest and most universal genre of representational art. Some are found in prehistoric caves. There are ancient specimens from Egypt, China, and Greece. Intricate eighteenth-century silhouette portraits are housed in London's National Portrait Gallery.

This simplest form of portraiture was a stepping stone to early photography. Instead of tracing a shadow by hand, John Herschel had nature fix the image with potassium ferricynanide in 1839. Botanists would place a plant directly on the photosensitive paper and shine a strong light on it. The blue and white image, a precursor of the blueprint, was sensitive to the varying opacity of the plant—unlike simple silhouettes. Cyanotype photographers were aware that the images were of the back sides of the plants; for they warn novices to turn the intricate features face down. Further piecemeal improvements in photochemistry and optics allowed the object's image to be taken without direct contact with the photosensitive paper. Further improvements decreased exposure time, increased the durability of the photograph, and formed a continuous developmental sequence leading up to modern photography.

---

[8] I owe this point to Laurie Paul. She notes that the qualification is not needed if Far is stationary

Night photography harkens back to these roots. Nazi naval archives contain ominous silhouette photographs of targeted ships. Although New York City was out of range of Nazi bombers, cargo ships using New York harbor were in range of Nazi submarines. The cargo ships turned off their sea lights but were back-lit by the bright lights of Manhattan. (This explains why New York City had to be blacked out during World War II.) The Nazi archives also include silhouette photographs of the effects of cargo ships—their wakes, smoke, debris, and the like. These photographs of ship effects can be distinguished from the photographs of the ships themselves. Consider a submarine captain who first spots traces of a ship against the moonlight. Since the captain *sees that* a ship is in the vicinity, he follows its trail until, finally, he *sees* the ship itself profiled against the moonlight.

Purely contrastive seeing is not an ability peculiar to *homo sapiens*. Aquatic predators frequently hunt from below. Light enters water from above, so prey are back-lit. A shark that approaches a school of anchovies from the side will be dazzled by the light reflected by the anchovies's scales as the school sharply pivots in unison. Some sharks counter by driving the school toward the surface and then lunging in from below. The sharks see the anchovy silhouettes. They also see the anchovies.

One rationale for counting the sighting of a silhouette as a sighting of the object might be that the visual match is sufficiently good—or as good as obtained during nocturnal conditions when we see only in black and white. This fidelity explains why silhouettes are so widely used as iconographic symbols.

As in the case of seeing transmitted objects, the good visual match holds only as a general trend. The silhouette of a pile of junk does not look much like a pile of junk. But if I see a silhouette of a pile of junk, I see a pile of junk. Interviewers preserve the anonymity of informants by televising them in silhouette. Viewers see the informant but cannot identify him on the basis of what they see.

Under some conditions, a cast shadow provides just as strong a visual match as the silhouette. Early silhouette portraitists traced the sitter's shadow as it was cast against white paper. Casting a shadow on a cloud with a powerful search light makes it seem as if one is seeing an object back-lit from the sky above. Hence the "bat signal" from the Batman comic strip is called a silhouette.

An object's genuine silhouette has a more robust resemblance to the object than the object's cast shadow. The cast shadow is only comparable to the corresponding silhouette under a narrow set of conditions—

and only if we ignore the spatial displacement. Causal theorists reject the hypothesis that seeing is a matter of having experiences that resemble the scene. Resemblance is neither a necessary nor a sufficient condition for seeing. Causal theorists agree that there is generally a good visual match, but seek to explain that fidelity in terms of causal connections between visual experience and the scene. *Why* do silhouettes generally provide a better match than cast shadows? Why do I see an object by seeing its silhouette but not by seeing its cast shadow?

Note that I am using 'see' in Fred Dretske's[9] sense of nonepistemic seeing. Whereas seeing *that a* is *F* entails belief that *a* is *F*, nonepistemic seeing lacks belief content (*ibid.*, p. 88). When cave men witnessed a solar eclipse, they saw the moon even if they had no beliefs about what they were seeing.

Or consider the original situation in which I am in doubt about whether I am seeing Near or Far. Given my lack of belief, I cannot be said to see *that* the silhouetted object is Near or see *that* the silhouetted object is Far; for I lack the belief that the object is Near and lack the belief that the object is Far.

## III. CAUSAL PROCESSES AND PSEUDO-PROCESSES

Wesley Salmon[10] draws a distinction that helps the causal theorist explain why seeing the silhouette of an object is seeing the object but seeing its cast shadow is merely seeing an effect of the object. When a blimp casts a shadow that races along the ground, the moving shadow has persistence conditions that are entirely parasitic on the blimp and the sun. I can predict where the shadow will move by keeping my eyes on the ground and tracking the shadow's speed and direction. But the stages of the shadow's movements are not related as cause and effect. Rather, they relate to the stages of the blimp's movements. Elliott Sober[11] diagrams the causal structure by using → to represent causal connections and broken arrows as the absence of a causal connection (*ibid.*, p. 303):

$$\begin{tikzcd}
\text{causal process} & C1 \arrow[rr] \arrow[dr] & & C2 \arrow[rr] \arrow[dr] & & C3 \arrow[rr] \arrow[dr] & & C4 \arrow[rr] \arrow[dr] & & C5 \\
\text{pseudo-process} & & P1 \arrow[rr, "/" marking] & & P2 \arrow[rr, "/" marking] & & P3 \arrow[rr, "/" marking] & & P4 &
\end{tikzcd}$$

---

[9] *Seeing and Knowing* (Chicago: University Press, 1969), p. 88. Do not picture nonepistemic seeing as restricted to agnostic viewing. Nonepistemic seeing is compatible with being heavily opinionated about what one is seeing. Dretske's point is that this sense of 'see' does not *entail* any beliefs.

[10] *Scientific Explanation and the Causal Structure of the World* (Princeton: University Press, 1984).

[11] "A Plea for Pseudo-Processes," *Pacific Philosophical Quarterly*, LXVI (1985): 303–09, here p. 303.

If I watch the moving shadow cast by the blimp, then I see a pseudo-process governed by the blimp's movement. But if I watch the silhouette of the blimp, I see a causal process. The same point holds if the blimp is stationary.

If the blimp collides with a skyscraper and is destroyed, the shadow is terminated. If only the shadow meets the skyscraper, the shadow is momentarily deformed but then, after passing the skyscraper, continues on exactly as before, carrying no trace of its encounter with skyscraper. I can enlarge the cast shadow by inserting a flag into the perimeter of the invisible shadow as the cast shadow passes in front of me. But this alteration does not persist after the shadow passes the flag. Putting a flag on the blimp does result in a lasting change. I can mark the blimp but cannot mark the shadow.

> A given process, whether it be causal or pseudo, has a certain degree of uniformity—we may say, somewhat loosely, that it exhibits a certain structure. The difference between a causal process and a pseudo-process, I am suggesting, is that the causal process transmits its own structure, while the pseudo-process does not. The distinction between processes that do and those that do not transmit their own structure is revealed by the mark criterion. If a process—a causal process—is transmitting its own structure, then it will be capable of transmitting certain modifications in that structure.[12]

The silhouette of the blimp is a causal process because it satisfies the mark criterion. A change of the silhouette is inherited by future stages of the silhouette.

Since an intrinsic change in the silhouette constitutes an intrinsic change in the object, the silhouette is part of the object. A silhouette, in this mereological sense, is not a depiction because it exists independently of any intention to represent the object.

The moon is like the blimp during a solar eclipse. The silhouette of the moon is a causal process. If the silhouette is changed by a meteor collision, then the silhouette retains that disfigurement. The shadow cast by the moon is a pseudo-process, however. As it races along the surface of the earth, local interventions have no effect on future stages.

The shadow moves as quickly as the earth spins. In principle, a shadow can move faster than the speed of light. Suppose I held my thumb over a flashlight that could transmit light across vast distances. I wave the flashlight, sending the shadow of my thumb coursing through the heavens. The rate at which the cast shadow moves increases geometrically with distance. At a large enough distance, my thumb's shadow is moving faster than the speed of light.

---

[12] Salmon, p. 144.

This is compatible with the special theory of relativity.[13] The theory says only that no *signal* can travel faster than the speed of light. The shadow is not a signal because it cannot transmit information. My thumb's shadow when cast on surface *A* does not pass on any information from *A* to any subsequent location *B*. The shadow's features at *B* instead depend on the flashlight and my thumb.

The silhouette of my thumb can move only as fast as my thumb; for the silhouette of my thumb is where my thumb is. The persistence conditions of a silhouette make it a causal process. A process is a persistence of something—a wave, a table, and so on.[14] Both the object and its silhouette are genuine causal processes.

Plato's allegory of the cave was inspired by the shadow plays of puppeteers. In the allegory, men are shackled together in a way that keeps them facing a cave wall. Behind and above them is a fire and a walled walkway. The barrier conceals servants who stroll by with figurines above their heads. These figurines cast shadows on the cave wall. This shadow play is the only reality for the prisoners who have never seen things under normal conditions. The prisoners consider the shadows as objects in their own right. They do not view them as signs of other objects. They do not know that the shadows are cast by figurines and so do not use the shadows to make inferences about those figurines. When the prisoners view the shadows in this fashion, they do not see the figurines. My explanation is that one can see only an object via a causal process emanating from that object. The prisoners do not see any objects even though they can predict the patterns of shadow play.

If a prisoner is turned toward the silhouettes of the figurines, then he sees the figurines—whether he recognizes them or not. Suppose the prisoner comes to have normal knowledge of the world but resumes his original station in front of the shadows. This prisoner knows that the figurines are casting the shadows. He can see *that* the present shadow is being cast by the horse figurine. Yet he no more sees the figurine than his ignorant companions. He sees exactly what his ignorant cavemates see. Since nonepistemic seeing has no belief content, what one sees cannot be increased by learning.

Astronomy teachers warn that staring at an eclipse causes irreversible eye damage. Galileo's telescope magnified this destructive power and contributed to his blindness. Hence teachers encourage students to view eclipses by means of pinhole cameras. The students can safely watch the shadow cast on the ground by the moon as it crosses in front of the sun. Teachers assure the students that they are seeing the moon eclipse the sun. But the students are actually like the enlightened cave prisoner who knows the shadow is cast by an object that is, sadly, out of view. Only a student who steals a glance at the silhouette in the sky sees the moon eclipsing the sun.

---

[13] Milton A. Rothman, "Things That Go Faster Than Light," *Scientific American*, CCIII, 1 (July 1960): 142-52.

[14] Salmon, p. 140.

## IV. SILHOUETTES AS ABSORPTION SURFACES

The silhouette is the *surface* of the object that makes it visible by virtue of the light that it blocks. Therefore, to see a silhouette of an object is to see a *part* of it. Seeing a relevant attached part of an object suffices for seeing the object *simpliciter*. The reason why we count seeing the silhouette of an object as seeing the object is that we are seeing a surface of the object.

Leonardo Da Vinci propounds a whittling objection to the assumption that the surface of an object is the outermost material part of an object.[15] If the surface is the *outermost* part of the object, then how can it have any thickness? If it were one millimeter thick, then only the outermost 1/2 millimeter could be the outermost part of the object. That 1/2 millimeter could not be the outermost part because only *its* outer half could qualify as the outermost part of the object. That outer 1/4 millimeter can itself be halved. And so on. Consequently, surfaces have 0 thickness and thus cannot be a material part of the object. Da Vinci goes on to infer that the surface is a common boundary between an object and its environment.

The whittling argument illicitly manipulates the domain of discourse for the quantifier 'outermost'. What counts as the parts of an object varies with purposes. Sometimes the parts include a coat of paint, the enamel, the peel, and so on. Sometimes not. The domain is nondissective; parts of the parts need not be in the domain of discourse. If I say that the surface of my couch is plaid, I am not refuted by the observation that none of the small parts of the surface is plaid. Da Vinci's whittling argument can proceed only by shifts of context to smaller and smaller parts. Thus, the argument that material objects lack surfaces is much like Peter Unger's[16] argument that no material object is flat. Both of these arguments are refuted by the kind of contextual analysis David Lewis proposes in "Scorekeeping in a Language Game."[17]

---

[15] See Avrum Stroll, *Surfaces* (Minneapolis: Minnesota UP, 1988), pp. 40-46.

[16] *Ignorance* (New York: Oxford, 1975)

[17] *Journal of Philosophical Logic*, VIII (1979): 339-59.

Like most philosophers and psychologists, I say that we see objects only by seeing their surfaces. Stroll[18] cites Venus as a counterexample: we see Venus even though the surface of Venus is completely covered with clouds. I say that we see Venus only if we count the clouds of Venus as part of Venus. If we count Venusian clouds as not part of Venus, then we do not see Venus. The same holds for objects that are marked to make them visible. If I paint part of a perfectly transparent ball black, then I see the ball only if the paint is part of the ball.

Stroll denies that Jupiter has any surface at all since it is gaseous. Here, I concede that it is vague where the surface is. But the vagueness of the location is compatible with Jupiter having a genuine surface. Microphysicists regard the surfaces of all objects as vague.

Stroll rightly observes that philosophers and perceptual psychologists have done too little to defend their presupposition that all objects have surfaces. And he is right again about their lack of support for the presupposition that we see by virtue of these surfaces. But I think both presuppositions are correct. Therefore, I rely on both presuppositions when assimilating purely contrastive seeing to the case of seeing objects that transmit light. But I endorse Stroll's objection to the assumption that we see by virtue of a single surface. Purely contrastive seeing shows that there are at least two types of surfaces that play a role in vision.

Surfaces have physical properties. In front-lit conditions, we see an object by virtue of the light transmitted by its front layer. In back-lit conditions, we see the object by virtue of light blocked by its back layer. A layer blocks by a different mechanism than it reflects. Consequently, the surface of a back-lit object consists of the portion of the object that is just enough to block the light. This absorptive surface must be distinguished from the reflective surface. We can represent the two surfaces of the far side of the moon like so: )). The outer parenthesis is the surface relative to light reflection. The inner parenthesis is the surface relative to light absorption. An astronaut in orbit over the far side of the moon during a solar eclipse does not see the silhouette of the moon even though he sees the reflective surface of the far side of the moon. Viewers from earth are seeing the absorption surface.

More precisely, the viewers from earth are looking into a concave, bowl-shaped surface. It is indistinguishable from a dark disk just as the silhouette of a cube (viewed corner on) is indistinguishable from the silhouette of a flat hexagram. But we are more apt to interpret the silhouette of the moon (in particular, the near side of the moon) three dimensionally as a *convex* surface. Hence the temptation to infer that the moon or sun has turned black. Most objects *are* convex rather than concave. Hence our strong preference for convexity could simply be due to the incorporation of an empirical regularity into our visual system.

---

[18] *Surfaces*, pp. 75-76.

Familiar seeing by transmitted light is surprisingly interpretative. The image of a square could be produced by an infinite variety of trapezoids or tilted rectangles. But silhouettes are much more interpretive because much less information is conveyed. Perceptual psychologists are impressed by how many of these physically possible interpretations never occur to the viewer. Usually, we promptly commit to a single interpretation. David Marr[19] concludes that we disambiguate by hard-wired guidelines:

> Somewhere, buried in the perceptual machinery that can interpret silhouettes as three-dimensional shapes, there must lie some source of additional information that constrains us to see silhouettes as we do. Probably,...these constraints are general rather than particular and do not require a priori knowledge of the viewed shape (*ibid.*, p. 219).

Marr articulates these assumptions as if the lines of sight emanate from an illumined object. If each point on the silhouette corresponds to one point on the viewed surface, we obtain a curve that serves as a contour generator. The three-dimensional shape of the object can be inferred, given Marr's further pair of assumptions that nearby points on the contour of the image correspond to nearby points on the object and that all points on the contour generator lie in a single plane.

Thus, Marr underwrites our tendency to view the silhouette of the moon as *convex* with this rational reconstruction. If he is correct, our visual system violates the causal theory of perception. The system assumes, in effect, that we are seeing a part of the object that is causally idle, namely, the front surface of a silhouetted object. This hard-wired heresy might contribute to the counterintuitiveness of my thesis that we see the backs of silhouetted objects.

Regardless of whether I am laboring against a neurological illusion, my thesis that we see the (inner) far side of the moon during an eclipse is paradoxical. The general reason is that we tend to model vision on cases where we see objects by means of the light they transmit. This makes it seem like we are seeing through the 3,600 kilometers of solid rock between us and the far side.

---

[19] *Vision* (San Francisco: Freeman, 1982).

To lessen the appearance of a miracle, some people suggest that I am only seeing the edge of the moon during a solar eclipse. The edge can block light, however, only if it is a physical object rather than an abstract mathematical limit. To be visible from earth, that edge would have to be many meters wide. If I can see the backside of this giant ring of matter, then I would still be seeing a surface that is behind several kilometers of solid rock! Reducing the quantitative scale of a miracle secures no advantage: "Don't be so incredulous; Jesus only restored sight to *one* eye of the blind man." Half a miracle is still a miracle.

Suppose a florescent balloon meanders in front of a solar eclipse. First the glowing balloon blocks our view of the bottom edge of the moon, then the middle of the moon. But if the balloon *blocks* our view of the middle of the moon, then we must have previously seen the middle of the moon. Therefore, we were not just seeing the edge of the moon during the solar eclipse.

Here is a summary of the basic argument for the thesis that I see the entire backside of silhouetted objects. When I see the silhouette of an object, I see the object. I see an object only if I see part of it. That part must cause my perception in an appropriate way. The only part that can play this causal role is the object's absorption layer. Therefore, when I see a silhouetted object, I see its back surface.

## V CONFUSION ABOUT OCCLUSION

Objects that are seen at least partially in virtue of the light they transmit can be blocked from view by interrupting their transmission of light. But a perfectly dark object does not owe its visibility to the light it transmits. Hence, these nontransmitters are an exception to the principle that one can always conceal an object by interposing an opaque body. We do not need to see through the "blocking" object.

This point solves a puzzle about cast shadows. The first version was devised by Robert Fogelin around 1967 or 1968. It became a topic of lunchtime conversations with his colleagues at Yale University: Charles B. Daniels, Robert Stalnaker, Richmond Thomason, Samuel Todes, and Bas van Fraassen.[20] The Yale puzzle and its solution were presented in detail by Todes and Daniels.[21] But I shall present the simplified version van Fraassen mentions in *Laws and Symmetry*.[22] Although 'shadow' is often used as a count noun, it should be read below as a mass noun (as in 'How much shadow is there in the photograph?'):

---

[20] More oral history. My chief informant, Fogelin, was dissatisfied with a colleague's analogy between 'Only I can feel my pain' and 'Only I can cast my shadow' Fogelin devised a complicated enigma involving infinitely many shadow casters. Fogelin is not sure whether the simplification to two shadow casters was his doing, the doing of another individual, or the product of the Yale collective consciousness.

[21] "Beyond the Doubt of a Shadow: A Phenomenological and Linguistic Analysis of Shadows," in Ihde and Zaner, eds., pp. 86-93.

I. If *X* casts any shadow, then some light is falling directly on *X*.
II. *X* cannot cast shadow through an opaque object.
III. All shadow is shadow of something (*ibid.*, p. 217).

Imagine a barn casting shadow on a sunny day. A bird flies between the barn and the shadow cast on the ground. The shadow directly beneath the bird cannot be cast by the bird (by virtue of I). Nor can it be cast by the barn (by virtue of II). But no third thing can cast the shadow. Hence, III is violated.

Principle II is true but insinuates that the only way that a shadow can appear on the far side of an intervening object is by penetration.[23] It discourages us from considering the alternative that the shadow appears on the far side by default. Nothing aside from the original blockage of light is needed to place shadow there.

It is true that (unreflected) light can appear on the far side of an intervening object only by penetrating it. But opaque bodies do not block shadows. Nor do they let shadows pass through. It is also a mistake to picture a transparent pane of glass as letting a shadow pass through. The glass only lets light pass through. But this does not mean that glass blocks darkness. Talk of blocking or transmitting shadows is a symptom of the fallacy of reification.

A linguistic philosopher might be tempted to trace the reification to the fact that 'shadow' is a mass noun even though it has no mass. I suspect a stronger influence, however, is the representational economy of treating shade as if it were colored light. This fiction lets us effortlessly extend the projective geometry that is used to depict illumined objects to the depiction of shadows. Artists sometimes draw with white chalk on dark paper. They represent shadows with undrawn regions. This omissive method of representing shadows echoes scientific reality. We normally use light paper, however. To depict a dark area, we must shade in with a pencil or some dark substance. This practice abets the reification of shadows.

Darkness is an absence of light, not a substantive force that can be stopped with shields. Once the light has been stopped, nothing further needs to be done to ensure darkness. This asymmetry of manipulation is a general difference between properties and their privations. A homeowner who insulates his house cannot literally keep the cold out; he can only keep the heat in. Cold is the privation of heat and hence is not a substance that can be directly manipulated.

---

[22] New York: Oxford, 1989.
[23] Todes and Daniels, p. 90.

Cold spots, shadows, and holes are not objects.[24] Therefore, they do not challenge the principle that perceivers must have a causal connection with the seen objects. To see a hole is to see a perforated object. It is not the same as seeing that the object is perforated because one might mistakenly believe that one of the perforations is a spot of dark paint. We nonepistemically see holes. But this just amounts to nonepistemically seeing an object that has a certain shape. When I see a dark doorway, I see a holed wall. Behind every privation is a substance that suffers the privation.

## VI DARK OBJECTS MUST EFFECT CONTRASTS

W. R. Brain[25] regards perception of perfectly dark objects as a counterexample to the causal theory of perception. But the causal theory correctly insists that we need light to see—if not light transmitted by the object, then light transmitted by the object's immediate surroundings.[26] Suppose my only source of light is a window that has two shutters. When the left shutter is closed, I see it by virtue of how its right side is outlined against the remaining light. When both shutters are closed, I see neither shutter.

The main reason for saying that the overshadowed smaller suitcase is invisible is that the smaller suitcase does not causally contribute to what is seen. The overshadowed suitcase also violates the requirement that *S* nonepistemically sees *O* only if *O* "is visually differentiated from its immediate environment by *S*." [27] Consider a white moth on a white tree trunk. The white moth causes part of what I see, although not in a way that enables me to discern its outline. I am looking at the moth but I am seeing it.

Now consider a perfectly dark moth on a perfectly dark tree trunk. This case differs from the cases involving back-lit occlusions. The dark moth does not have a silhouette. The scene is front lit. Indeed, I am shining a strong light right on the moth. The moth is absorbing all the photons that land on it and hence is playing a causal role in what I am seeing. For instance, if the dark moth is atop a white spot, the moth prevents me from seeing that white spot. Thus, the dark moth satisfies the causal theorist's necessary condition for seeing. The moth is not an idler like Near.

---

[24] The temptation to reify holes can be strengthened with seductive arguments. These sophisms are beautifully illustrated and lovingly examined in Roberto Casati and Achille Varzi's *Holes and Other Superficialities* (Cambridge: MIT, 1994).

[25] "The Neurological Approach," in R. J. Hirst, *Perception and the Physical World* (New York: Macmillan, 1965).

[26] Richard J. Hall, "Seeing Perfectly Dark Things and the Causal Conditions of Seeing," *Theoria*, III (1979): 127-34.

[27] Dretske, p. 20. For criticism of the visual differentiation condition, see chapter four of Gerald Vision's *Problems with Vision* (New York: Oxford, 1997).

When the dark moth is on a white tree trunk, the moth is seen by the contrast it creates via its light absorption. Absorption is a well-understood physical mechanism as evident from black-body research in physics. True, the white bark of the tree also contributes by reflecting light. But seen objects need not make the *entire* causal contribution.

Transmission of light is not a sufficient condition for being seen. The stars shine as brightly during the day as at night. If you put the halves of a ping-pong ball over your eyes, you do not see any objects. Transmitters rely just as much on contrast as dark objects. Human beings have exquisite contrast sensitivity that strains the capabilities of electronic display devices, such as television and computer graphics displays. People can detect a change of modulation of less than one percent across a border.[28] This extreme sensitivity to differences in light intensity explains how the "darkness" of an object can itself be a pure contrast effect. Sunspots look dark solely because they are relatively cool spots on the sun. They would shine brightly if they could be removed from the even hotter photosphere of sun. The universality of this reliance on contrast explains why any object can be camouflaged by muting its contrast with its surroundings. Seen objects normally make a *holistic* contribution to the scene. They are visible by virtue of the differences that they make with their surroundings. The differences can be achieved by going up or down the scale of light intensity—including all the way down to zero.

The contrast between the silhouette and its background is muted by semi-opaque intermediates, such as window curtains. Since our eyelids are only semi-opaque, it is possible to see some objects with our eyes closed. Close your eyes and position a bright light bulb so close to them that you can feel the bulb's heat. Now wave a pen just in front of your eyelids. (Contrast sensitivity is enhanced by movement.) Seeing the pen's diffuse silhouette may make you receptive to the hypothesis that purely contrast seeing is biologically prior (phylogenetically and perhaps ontogenetically) to seeing objects by their transmitted light.

---

[28] Robert Shapley and Dominic Man-Kit Lam, *Contrast Sensitivity* (Cambridge: MIT, 1993), p. xii.

All evolutionary accounts start with a proto-eye that is merely sensitive to the presence of light and then develops into a detector of passing shadows. The first seen objects should have been silhouetted figures. Only later would we expect objects to be seen by virtue of the light they transmit. Instead of being a marginal form of vision, purely contrastive seeing is the primal form of seeing. Stereotypical seeing is an elaboration of this more basic ability.

Black holes amass the immediate surroundings that make themselves discernible. The textbook scenario features a black hole that develops from a large star in a binary star system. The larger star has been pulling matter from the smaller star. After the large star accumulates a critical amount of matter, it collapses into a black hole. Since the black hole has the same mass as it had when a large star, it continues to pull matter from its partner star. Since this new matter conserves angular momentum, it forms an accretion disk that revolves around the black hole. The matter from the inner part of the disk accelerates up to nearly the speed of light. The friction creates tremendous heat which in turn emits X-rays. Hence the black hole makes itself visible by piecing together an environment for itself.

Goldman rightly denies that there is a "transfer of energy or force" from the black hole to the observer (*ibid.*, p. 282). But he wrongly denies that there is a physical mechanism by the which the black hole makes itself seen. The black hole causes other things to transfer energy to observers. During World War II bombing raids, high-flying bombers could see the silhouettes of low-flying bombers below: ✈. The low bombers were back-lit from enemy searchlights. The low-flying bombers were not directly transferring energy into the eyes of high-flying observers. But the low-flying bombers were still causing themselves to be visible by rousing the enemy below.

Dark objects are a varied lot. Nevertheless, neither the overshadowed suitcase nor the overshadowed heavenly body, Near, satisfy the causal requirement. True, I would see Near if Far vanished. But I see in virtue of actual causation, not hypothetical causation. Recall Harry Frankfurt's[29] counterexample to the principle that I am responsible only if I could have done otherwise. An evil scientist has rigged up a device that will make me do an evil deed if and only if I fail to do it by my own accord. I do the evil deed, so the device does nothing. I am responsible for the evil deed even though I could not have done otherwise.

---

[29] "Alternate Possibilities and Moral Responsibility," this JOURNAL, LXVI, 23 (December 4, 1969): 829-39.

Compare Frankfurt's case with Lewis's[30] censor. The censor is a device that will make me have the visual experience of a scene if and only if I do not have the visual experience in another way. I am having a visual experience of the scene by opening my eyes and receiving reflected light waves in the normal way. Am I seeing the scene?

Lewis says I see if and only if "the scene before my eyes causes matching visual experience as part of a suitable pattern of counterfactual dependence" (*ibid.*, p. 285). The censor illustrates a lack of suitable dependence:

> The case is one of causal preemption. The scene before my eyes is the actual cause of my visual experience: the censor is an alternative potential cause of the same effect. The actual cause preempts the potential cause, stopping the alternative causal chain that would have otherwise gone to completion (*ibid.*, pp. 285-86).

Substitute 'Near' for 'the censor'. Although the censor is idle, its "idleness is an essential factor in the causal process by which matching visual experience is produced.... We cannot uniformly ignore or hold fixed those causal factors which are absences of intervention" (*ibid.*, p. 286). Lewis concludes that I do not see the scene.

Lewis's verdict has been unpopular. His personal envoy, Bruce Le Catt[31] has tinkered with the counterfactual dependency theory to illuminate the possibility of securing a better match with intuition. But these changes do not address the main problem: underweighting the importance of physical mechanisms. As Brian McLaughlin[32] stresses, I see because the censor is an external fail-safe device that does not prevent me from exercising my capacity to see.

Let me turn the topic temporarily from nonepistemic seeing to perceptual knowledge. The eclipse riddle does not fit into the tradition of skeptical counterpossibilities. It is not like cases involving dreams, brains in vats, or fake barns.[33] I have access to all the relevant facts. When I observe Near and Far approach and then intersect, I know they are behaving just as astronomers predicted many years ago. My situation differs from that of a passerby who had only heard of Near and just happens to look up at the moment of intersection. He might be surprised to learn that the dark shape might be caused by Far. Not me; I have been preparing for the event for years and know all the empirically relevant facts.

---

[30] "Veridical Hallucination and Prosthetic Vision," in *Philosophical Papers*, Volume 2 (New York: Oxford, 1986), pp. 273-86; here p. 285. Originally published in *Australasian Journal of Philosophy*, LVIII (1980): 239-49.

[31] "Censored Vision," *Australasian Journal of Philosophy*, LX (1982): 158-62.

[32] "Lewis on What Distinguishes Perception from Hallucination," in Kathleen Akins, ed., *Perception* (New York: Oxford, 1996), pp. 198-231.

[33] Goldman, "Discrimination and Perceptual Knowledge," this JOURNAL, LXXIII, 20 (November 18, 1976): 771-91.

The only threat to my perceptual knowledge of Far is conceptual unclarity. If I fail to believe that I am seeing Far, then I do not know that I am seeing Far. Astronomers saw dark nebulae long before they had perceptual knowledge that dark nebulae are clouds of interstellar dust. They had surprising difficulty excluding the possibility that the dark regions in the sky are tunnels through which we look out beyond the stars of the Milky Way into intergalactic space.

## VII. CAUSE, RESEMBLANCE, AND VISUAL DETECTION

We are tempted to say we see Near because we seem to see its outline (and so appear to satisfy Dretske's condition of visual differentiation). The influence of the outline can be gauged by considering a scenario that increases the relative size of Far. The huge shadow would make Near invisible. Those tracking the movement of Near would say that they had lost sight of Near as it is submerged into the massive umbra of Far.

The extent of overshadowing is irrelevant to occlusion as long as the entire object is covered. But the size of the shadow is relevant to visual detection. A thief who hides in a shadow that perfectly coincides with his shape is not as well concealed as one who enjoys coverage by a massive shadow that is not congruent with his shape.

When Near and Far have their original dimensions, the trackers of Near would be able to see that Near is aligned with Far even though they cannot see Near. To see an object, the object must be causally responsible for the visual information. This requirement explains why I do not see Mario Andretti when looking at his identical twin Aldo Andretti. A photograph of Aldo Andretti is not a photograph of Mario Andretti, even if their resemblance is exact. The astronomers are seeing an object that has the same look as Near. This resemblance ensures that what they see gives them all the visual information about Near that would have been available had Near caused the image itself. Hence, the astronomers tracking Near are reliably guided by the resemblance just as a reporter tracking Mario Andretti is reliably guided by a photograph of his identical twin Aldo Andretti.

Visual detection does not suffice for seeing. Astronomers first visually detected planets outside our solar system by observing perturbations of stars through ground telescopes. Only later, with the help of the Hubble telescope, did they see a planet outside our solar system.

## VIII THE COLLAPSE OF THE COMPOSITE ANSWER

The Far answer is unaffected by the third possibility that I am seeing a composite object whose parts are Near and Far. This mereological sum, Near+Far, is arbitrary but no more arbitrary than scattered objects, such as constellations. We do speak of seeing the moon even when we do not see the far side of moon. Seeing a relevant, properly attached part of an object counts as seeing the object simpliciter.[34] Seeing behaves just like most transitive verbs in this respect; to scratch a relevant attached part of an object is to scratch the object itself. Near is the front of Near+Far and Far is its back. If I see the front of Near+Far, then I see Near. But then the simple answer, that I see Near, would also be correct. The appeal to mereological sums would also be otiose if the simple 'I see Far' answer is correct. For if I see the back of Near+Far, I see Far. The composite answer is correct only if one of the simple answers is correct.

A polarizing light filter casts no shadow. But if a second filter of opposite polarity is placed beneath, then the pair jointly cast a shadow. Someone viewing an eclipse involving two such filters might qualify as seeing their mereological sum. Both the near filter and the far filter would each contribute to what is seen. In contrast, the heavenly body Near is completely idle. The same phenomenon arises for some objects that reflect light. Glass mirrors also polarize light. Hence glass mirrors darken when viewed through a polarizing filter even though everything else in the scene looks perfectly ordinary. The dark mirrors are visible by virtue of their contrast with their surroundings. But unlike Near, the mirrors causally contribute to the scene.

## IX VISUAL DIFFERENTIATION AND THE FAR SIDE

Picture Near and Far coming closer together until they fuse. I still see Far. This reinforces the point that we see the *backs* of back-lit dark objects. When the moon eclipses the sun, only the far side of the moon blocks the light. The front half is in the shade. Thus we see the far side of the moon during a solar eclipse.

True, little of the far side can be differentiated by purely contrastive seeing. Only the outer rim offers positive detail. The middle only supplies the negative information that there are no tunnels running completely through the moon. The silhouette of a fine mesh screen has a much higher proportion of positively detailed surface.

We distinguish between observing an effect of an object and seeing the object. The star Algol, in Perseus, has a companion star that is too dim to be seen. Since the plane on which these two stars revolve is oriented nearly edge on to our line of sight, each star is eclipsed by the other during every revolution. Although the effect of Algol's companion can be discerned by how it periodically dims Algol's light, astronomers describe Algol's companion as invisible. Astronomers cannot even discern its outline. The object must look some way to the perceiver.[35]

---

[34] McLauglin, "Perception, Causation, and Supervenience," *Midwest Studies in Philosophy*, IX (1984): 569-92, here pp. 580-85.

The analogy with Algol's companion suggests that during a solar eclipse, we see an effect of the moon but not the moon. The distinction between seeing and seeing effects can also be drawn in another way that appears to yield the common-sense answer that we see the moon but not the far side of moon. "Seeing the search beam of a lighthouse may count as seeing the lighthouse and not as seeing the bulb inside the searchlight."[36] The suggestion is that the moon stands to its far side as the lighthouse stands to its bulb.

The far side is not as obscured as the bulb, however. Aristotle argued that earth must be round because it always casts a round shadow on the moon during lunar eclipses. We have doubts about whether we see the earth this way because its shadow is displaced; the earth's shadow is on the surface of the moon but the earth is not. Similar reservations affect seeing objects by mirror reflection. But during a solar eclipse, the silhouette of the moon is not displaced. The information we get from the moon's silhouette is normally "discounted" by our ample background knowledge of the moon. The empirical content of the silhouetted moon is salient, however, when the observers have unusual background beliefs. Suppose we are worried that the moon was struck by a giant meteor during the daytime when it is invisible. Instead of waiting until night to find out, we fortunately have the opportunity to check the moon's condition by observing an afternoon eclipse: "What a relief! There the moon is, right on schedule, just where it belongs, entirely intact. The meteor must have missed!"

Actually, the detail from the far side during an eclipse can be greater than that offered by the moon's reflected light during a hazy evening. In an 1836 eclipse, Francis Baily noticed some bead-like features along one side of the moon. "Baily's beads" are the peaks of mountains that lie along the profile of the moon. If the moon revolved with respect to Baily and if the eclipse lasted long enough to complete one revolution, Baily could have surveyed all of the moon's peaks. But the moon always presents the same side to us. Why? Because the near side bulges about .5 kilometers and so is pulled more strongly by the earth's gravity. If a meteor had knocked off this bulgy excess, the moon might have been sent spinning fast enough to make silhouette astronomy informative.

---

[35] Dretske, p. 20.

[36] "Perception, Causation, and Supervenience," p. 584.

Granted, even Baily's hypothetical survey hardly compares to the rich visual information relayed by the spacecraft, Luna III, in 1959. But even the proud Soviet astronomers granted that, under certain conditions, we very dimly see a portion of the "dark" side of the moon by earthlight (a phenomenon popularly known as "the old moon in the new moon's arms").

When I see the moon in normal conditions at night, I see the moon by virtue of seeing its front. Sticklers deny that I see the moon; I see only the front surface. The same sticklers would say that on my account, we do not see the moon when it eclipses the sun; we see only the back surface. I admire the insight and attention to detail that motivates the stickler. But I am not a stickler. I say we see the moon during normal conditions at night and that we see the moon during an eclipse. But I need insist only that the stickler's standards be applied uniformly to both cases.

Suppose Near and Far have the same intrinsic diameter. Since Far is further away, it then only prevents light from striking the middle of Near. The stickler will say that I see a "doughnut" composed of an outer ring of Near and an inner core of Far. It is doubtful, however, that I see Far in this circumstance. I cannot differentiate Far from its immediate environment. Do I see a little star if it is stationed in front of the sun? If the two are equally luminescent, I cannot differentiate the little star from its immediate environment. Suppose a wall has beige wallpaper. If someone neatly pastes a patch of leftover beige wallpaper on the wall, do I see it? Following Dretske (*op. cit.*, p. 23), I answer no.

I can clearly discern the outline of Near. True, I do not see its center because it is blocked by Far. But the block is not the sort that makes a difference to the visual match with the scene; if Far were absent, the scene would look the same. Since I am a nonstickler, I say that in that circumstance, I see Near because I see a relevant, attached part of it.

Parallel, duplicate blockers are unexpectedly common in our carpentered world of rectangular volumes. Consider a silhouetted box. Although its front side has the same dimensions as its back side, the front side is farther from the light source. Hence the shadow generated by the back side fails to include the outermost edge of the front. Consequently, perimeter of the silhouette is actually caused by the front side of the box. The stickler concludes we see a rectangular doughnut. But since the outline of the back side is invisible, the back side fails to satisfy the visual differentiation condition. I conclude we see the front of the silhouetted box. (More precisely, the relevant absorption layer starts from the inside of the front side.) Boxy shapes are rare in nature. Nevertheless, they constitute an important exception to the rule that we see the backs of silhouetted objects.

If I am standing with my nose against a concrete wall, then I see the wall or at least a portion of it. Yet I am not visually differentiating the wall from its immediate environment. Dretske treats this anomaly as a limiting case (*op. cit.*, p. 26). He reasons that since there is no ***immediate*** environment, the requirement of visual differentiation does not apply. Interestingly, there is no parallel anomaly for the visual differentiation requirement in the case of purely contrastive seeing. If I am hurtling toward the moon during a solar eclipse, the black disk grows bigger and bigger, and then eventually overflows my visual field. There is only darkness. I know the moon is directly ahead of me. But I do not see it. For I need light to see.

## X. THE CAUSAL THEORY'S RECOVERY OF PRIMAL VISION

I now summarize my main theses. Instead of adjusting principles to fit the intuition that I see Near, we ought to disown this piece of common sense as the effect of an over-generalization. The causal theory of perception unexpectedly but correctly rules out the possibility that we see Near. Its support for Far forces us to recognize that, typically, we see the backs of dark, back-lit objects, not their fronts. So 'occlude' must be relativized. That forces a corresponding relativization of 'field of vision'. Thus the causal theory of perception brings an ancient species of seeing to the surface. Purely contrastive seeing is a living fossil that has everyday seeing as a descendent. Some minor adjustments are needed to accommodate this recovered form of seeing. Our conception of the physical mechanisms that support vision must be broadened to encompass the various kinds of visible dark objects. To include back-lit objects, we must allow partial light blockage to count. Objects are ***not*** seen simply by virtue of a contrast with their environment. They must cause the contrast. That is the difference between Near and Far.

Intersecting eclipses can form from a fortuitous convergence of heavenly bodies. But they also form in a cyclical fashion. As the moon orbits earth, the earth eclipses the sun. The scene would be reminiscent of the famous photograph "Earthrise" taken from the vantage point of the moon. Now take a giant step back from moon by imagining how the scene looks to an astronaut in a high orbit around the moon. The earth is then aligned with the moon and the astronaut has both of them in his line of sight. Near is the moon. Far is the earth.

ROY SORENSEN

New York University

## BOOK REVIEWS

*The Logic of Provability.* GEORGE BOOLOS. New York: Cambridge University Press, 1993. xxxvi + 275 p. Cloth $69.95, paper $24.95.

This is a revised and expanded version of the late George Boolos's *The Unprovability of Consistency: An Essay in Modal Logic.*[1] Like its predecessor, it exploits very well the opportunity to present a very attractive little theory in modal logic in terms accessible to a wide class of readers with some knowledge of first-order predicate logic. (Chapters thirteen and fourteen demand a bit more, but can be skipped without other loss.)

A thirty-six page introduction offers a survey of the content of the book. For another excellent summary, in the context of some history of the field and especially of the author's contribution to it, see the Afterward, written by John Burgess, in the recently-published collection of Boolos's essays.[2]

The first twelve chapters, covering roughly the subject matter of the earlier book, along with chapters thirteen and fourteen, concern certain systems of *modal propositional logic* (MPL). The formulas of MPL are built up from *sentence letters* $p$, $q$, $r$, and the like, and the constant $\perp$ (*absurdity*) using the connectives $\rightarrow$ (implication) and $\Box$ (necessity). For example, the negation $\neg A$ can be taken as an abbreviation for $A \rightarrow \perp$. A system of MPL is called *normal* if its set of theorems is closed under modus ponens, necessitation (that is, if $A$ is a theorem, then so is $\Box A$) and substitution of arbitrary formulas for sentence letters, and includes all tautologies and all instances of the *distributive laws* $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$. The least normal system is denoted by $K$. Chapter one is devoted to the relative strengths of some natural normal systems.

Chapter four develops Kripke semantics for normal systems. A *frame* is a pair $<W, R>$, where $W$ is a set (of "possible worlds") and $R$ is a binary relation on $W$. ($wRw'$ means that, from the perspective of world $w$, world $w'$ is possible.) A *model* $M$ consists of a frame $<W, R>$ together with a relation $V$ between worlds in $W$ and sentence letters. Relative to $M$, the notion that a formula of MPL is true in a world $w \in W$ is defined as follows: (i) the sentence letter $p$ is *true* in $w$ if and

---

[1] New York: Cambridge, 1979.

[2] *Logic, Logic, and Logic: The Collected Papers of George Boolos*, with Introductions and Afterward by John Burgess, Richard Jeffrey, ed. (Cambridge: Harvard, 1998).

0022-362X/99/9601/50–53

only if $wVp$. (ii) $\bot$ is not true in $w$. (iii) $A \rightarrow B$ is true in $w$ if and only if either $A$ is not true in $w$ or $B$ is true in $w$. (iv) $\Box A$ is true in $w$ if and only if $a$ is true in $w'$ for every $w'$ such that $wRw'$. Call $A$ *valid* in $M$ if and only if it is true in every world relative to $M$, call it valid in a given frame if and only if it is valid in every model with that frame, and call it simply valid if and only if it is valid in all frames. Given a model $M$ and a world $w$ of $M$, if the premises of modus ponens are true in $w$, then so is the conclusion. If $A$ is true in every world in $M$, then so is $\Box A$ and, if $A$ is valid in a frame, then so is each of its substitution instances. In particular, every theorem of $K$ is valid. Many of the standard normal systems, initially picked out because they corresponded to some intuition about what "necessity" might mean, turn out to be characterized by some simple property of the relation $R$, in the sense that their theorems are precisely the formulas of MPL which are valid in all frames (called the frames *appropriate to* the system in question) in which the relation $R$ has this property (chapter five). For example, corresponding to the system $T$ obtained from $K$ by adding as axioms all instances of $\Box A \rightarrow A$ is the property that $R$ should be reflexive on $W$. Chapter six gives a construction of a *canonical model* for each consistent normal system $L$, that is, a model in which the valid formulas are *precisely* the theorems of $L$.

In provability logic, $\Box A$ is intended to mean that the formula $A$ "is provable." To make this precise, a particular formal system must be specified in which it is to be proved. The simplest case, treated in the most detail, is the system of so-called *Peano Arithmetic* (PA). If the realization* assigns to each atomic formula letter $p$ of MPL a sentence $p^*$ of PA, it extends uniquely to all formulas of MPL by taking $\bot^*$ to be $\bot$ itself, $(A \rightarrow B)^*$ to be $A^* \rightarrow B^*$ and $(\Box A)^*$ to be $Bew$ $(\ulcorner \mathrm{A}^* \urcorner)$, where $\ulcorner B \urcorner$ is the Gödel number of $B$ and $Bew(x)$ is the standard arithmetical formula expressing that $x$ is the Gödel number of a formula which is provable in PA. To use Burgess's terminology, we call a formula $A$ of MPL *apodictic* if and only if $A^*$ is provable in PA for all * and we call it *veracious* if and only if $A^*$ is true for all *.

The necessary information about PA and its provability predicate is outlined in chapter two and some elementary facts about the interpretations * are in chapter three. In particular, it is shown there that every theorem of the normal modal system (GL, named for Kurt Gödel and M. H. Löb) is apodictic and that every theorem of the normal modal system (GLS, 'S' for Robert Solovay) is veracious. Chapter nine contains Solovay's proof that the apodictic formulas and the veracious formulas are precisely the theorems of GL and GLS, respectively. It uses the fact (proved in chapter five) that the frames appropriate to GL are those in which $R$ is transitive and there are no infinite sequences $w_1\ Rw_2\ Rw_3$.... There are analogous results involving other normal systems of provability logic, again due to Solovay, where, instead of PA, the proof predicate is for Zermelo-Fraenkel set theory (chapter thirteen) and second-order number theory—so-called 'analysis' (chapter fourteen).

Chapter eight contains three proofs of the *Fixed Point Theorem*, which expresses in modal logical terms the slight-of-hand aspect of Gödel's theory of undecidable sentences: let $A(p)$ be a formula of MPL and assume that all occurences of the letter $p$ in $A(p)$ are within the scope of $\Box$—so $A(p)$ says nothing directly about $p$, only about the provability of sentences containing it. Then we can effectively find a formula $B$ of MPL containing no sentence letter that is not in $A$ and not containing $p$ such that, writing $\boxdot A$ for $\Box A \cap A$:

$$\boxdot\ (p \leftrightarrow A(p)) \leftrightarrow \boxdot\ (p \leftrightarrow B)$$

is a theorem of GL. For example, one algorithm for constructing $B$ from $A$, together with Solovay's theorem, yields, with $S$ any sentence of PA and $\vdash$ meaning provable in PA:

$$\vdash\ S \leftrightarrow \neg\, \mathit{Bew}(\ulcorner S \urcorner) \text{ if and only if } \vdash S \leftrightarrow \neg\, \mathit{Bew}(\ulcorner \bot \urcorner)$$

and:

$$\vdash S \leftrightarrow \mathit{Bew}(\ulcorner S \urcorner) \text{ if and only if } \vdash S \leftrightarrow \neg\, \bot$$

Chapters fifteen and sixteen discuss *bimodal* logic, obtained by introducing a second modal operator $\boxed{1}$. The operation * assigning sentences of PA to sentence letters is extended to all bimodal formulas by taking $(\boxed{1}A)^* = \omega\mathit{Bew}(\ulcorner A^* \urcorner)$, where $\omega\mathit{Bew}(x)$ expresses the property of being provable in PA with one application of the $\omega$-rule. Again, there is a system GLB of bimodal logic whose theorems are precisely the apodictic formulas, although the proof in this case cannot proceed from a simple Kripke semantics for the system.

The final two chapters, seventeen-eighteen, treat quantified provability logic. Here, instead of the formulas of MPL, we have the formulas of QML, built up as usual in first-order predicate logic, except that we admit, given the formula $A$, also the formula $\Box A$. As in the propositional case, we can introduce the notion of a realization * for this system, namely, an assignment to each predicate letter $P$ of $n$ variables of a formula $P^*$ all of whose free variables are among the first $n$ in a fixed list of the free variables. * is uniquely extended to assign to each formula $A$ of QML a formula $A^*$ of PL with the same free variables. In this case, however, neither the class of apodictic formulas nor the class of veracious formulas forms a finitely (or recursively enumerably) axiomatized system.

I used the book as text for a ten-week course on provability (propositional) logic for a class of both graduate philosophy students and undergraduates, some but not all of whom had already had a course covering Gödel's incompleteness theorems. The book was at good level for them, although the students who had not had previous exposure to the arithmetization of syntax found chapter two hard going. There is a sprinkling of good exercises, but they probably ought to be supplemented in such a course.

A list of typos: page xxiii, line 27, "n" for "**n**" and on line, ")$x$" for "$x$)"; page 12, Theorem 22, $\neg\bot$ for the first occurence of $\bot$; page 14, line 6, delete "by Theorem 21 (with $\bot$ for $p$)," line 7, "$\bot$" for the last occurence of "$p$," line 8, "$\bot$" for the last two occurences of "$p$," lines 9 and 10,"$\bot$" for "$p$"; page 19, line 1, replace "(non-logical axioms)" by "(non-logical) axioms"; page 23, line 5, "$F(x)$" for "$Fx$"; page 31, line 26, "$\imath$" for "i"; page 44, line 6, "$y_{lh(y)-1}$" for "$s_{lh(y)-1}$"; page 45, line 7, "$y$" for "$var(y)$"; page 47, lines 35-36, the case in which $F$ is the negation of one of these equations also has to be treated; page 51, line 13, "realizations" for "interpretations"; page 63, line 14, "inconsistent" for "1-consistent."

W. W. TAIT

University of Chicago

## MEMO TO AUTHORS

The editors cannot afford to send to the printer manuscripts that are difficult for the compositor to read or footnotes that need to be completed in proof. They ask that manuscripts submitted be prepared according to the following guidelines:

(1) All material, including the quotations and footnotes—especially footnotes—must be typed, double-spaced, on one side of the paper, with ample margins left and right, top and bottom.
(2) No bibliographies will be published; bibliographical material should be put into the footnotes.
(3) Footnotes should be complete, according to Journal style (see any issue). Journal citations should include name of journal in full, volume number, issue number, month or season, year, and pages. Book references should include publisher, city of publication, and date. Footnotes should be gathered at the end of the article.
(4) Quotations should be checked by the author; capitalization, italics, spelling, and the like should match the original. Corrections should not have to be made (expensively) in proof.
(5) If there are diagrams, they should be provided camera-ready by the author. If there are special symbols, they should be widely spaced and clearly marked; if there are many, a list should be provided. Standard logic symbols are available, and letters of the Greek and script alphabets; no boldface.
(6) Computer copy can be extremely difficult to read; only letter-quality is acceptable. It should be double-spaced and not justified. The print should be dark enough and the letters widely enough spaced so that both letters and punctuation marks can be read easily.

The editors are willing to read and evaluate manuscripts that do not meet these standards, provided they are genuinely legible and not unreasonably long. They may ask for revision of excessively long manuscripts, and they require reformatting of accepted manuscripts, if necessary, before scheduling for publication.

The *Journal of Philosophy* is published monthly, in issues that (ordinarily) contain only 56 pages of copy; this limits the length of papers the editors can normally publish to no more than 7,500 words, or eighteen Journal pages. A more precise way of determining length is by characters: 2,850 characters per Journal page, or 51,300 characters per Journal article of 18 pages.

The *Journal of Philosophy* does not publish papers that are primarily historical or expository; the editors urge that such manuscripts be submitted elsewhere.

THE JOURNAL OF PHILOSOPHY

VOLUME XCVI, NO. 2, FEBRUARY 1999

UNIVERSITY LIBRARY
AT PUR
CAMPUS
3-1-2000

# SECOND-HAND MORAL KNOWLEDGE*

Trust enters into the making of a virtuous person in at least two ways. First, unless a child has a sufficiently trusting relationship with at least one adult, it is doubtful that she will be able to become the kind of person who can form ethically responsible relationships with others. Infant trust, as Annette Baier[1] has reminded us, is the foundation on which future trust relationships will be built; and when such trust is irreparably shaken, the adult into whom the child grows may be forever cut off from intimacy. Second, a moral beginner must trust other people's moral judgment while she learns to be good. A child does not learn how to be good on her own. She follows the precepts and examples of others, absorbs the moral values of the community into which she has been born, and relies on others who, with luck, will not lead her astray but will turn out to be trustworthy in offering her guidance and correction. While there is some empirical evidence that children learn about morality better when given reasons in support of what they are being taught, part of what a child is being taught is what *counts* as a good reason. Thus, no moral beginner could be in a position to assess adequately the cogency of the reasons she is given until she has progressed some distance in her education.[2] Until that time, the beginner must rely on other people's judgment.

---

* This paper was written while I was a fellow at Cornell University's Society for the Humanities; I would like to thank the Society for its support. The paper was presented at *Different Voices: A Conference in Honor of Annette Baier*, held at the University of Pittsburgh in 1995, and to audiences at Cornell, the University of Cincinnati, and the 1996 Australasian Association of Philosophy Annual Conference—each audience contributed to the final shape of the paper Special thanks to Jennifer Whiting, Lisa Rivera, and Donald Ainslie for written comments on earlier versions.

[1] "Trust and Anti-trust," *Ethics*, xcvi (1986) 231-60.

[2] Some accounts of moral education (for example, Aristotle's) will give a greater role to trust than others do

0022-362X/99/9602/55–78

That trust enters into the making of a morally good person should thus be uncontroversial. I want to defend a much more controversial thesis: needing to trust other people's moral judgment is not just a stage we go through on the way to becoming morally virtuous. If the world of value is complex, and if our access to it is shaped by our experiences, then even among the morally mature there will continue to be a significant role for moral testimony and thus for trust. We should not think of our reliance on other people's moral judgment as like training wheels for the novice cyclist, wheels which can and ought to be dispensed with, once we have the know-how and the confidence to ride off on our own. Often, we cannot do as well on our own as we could do if we accepted the moral testimony of others. Moreover, just as borrowing scientific knowledge can enhance our capacity to discover truths about the nonmoral world, borrowing moral knowledge can enhance our capacity to understand the world of value.[3]

I am going to start by giving a sketch of what I take to be the best arguments *against* testimony as a source of moral knowledge. Next, I give a detailed real-life example which aims to make us suspicious of the case against testimony. At the same time, the example tries to be fair to the considerations against testimony and to show why we might find morally attractive someone who resists such borrowing and refuses such trusting. Armed with the example, I then return to reexamine the initial case against moral testimony. In the final section, I argue that making room for moral testimony brings issues of trust to center stage. The wise truster of moral testimony begins from a default stance of distrust, yet can come to trust and, through trusting, can come to have moral knowledge that would otherwise be inaccessible to her.

## I. THE CASE AGAINST TESTIMONY IN MORAL KNOWLEDGE

It has seemed simply obvious to some, such as Robert Paul Wolff,[4] that there is no place for testimony about moral matters:

---

[3] Throughout this paper, I am going to speak of "moral knowledge" and sometimes even of "moral facts," but this is simply for convenience of expression. While I do think that only a realist account of morality can adequately accommodate our moral practice, it is no part of my brief to argue for that here. Quasi-realists (for example, Simon Blackburn, in *Spreading the Word* (New York: Oxford, 1984)) and norm expressivists (for example, Allan Gibbard, in *Wise Choices, Apt Feelings* (Cambridge: Harvard, 1990)) can each offer nonrealist reconstructions of this sort of talk. Only on a very simple subjectivist view will such talk be illicit. And I take it—admittedly without argument—that such a view is no longer a contender.

[4] *In Defense of Anarchism* (New York: Harper, 1970); also cited in Gerald Dworkin, *The Theory and Practice of Autonomy* (New York: Cambridge, 1988), p. 52. John Hardwig shows how Wolff is wrong about contemporary mathematics, given the proliferation of mathematical specialties; see his "The Role of Trust in Knowledge," this JOURNAL, LXXXVIII, 12 (December 1991): 693-708.

> He [the responsible man] may learn from others about his moral obligations, but only in the sense that a mathematician learns from other mathematicians—namely by hearing from them arguments whose validity he recognizes even though he did not think of them himself. He does not learn in the sense that one learns from an explorer, by accepting as true his accounts of things one cannot see for oneself (*ibid.*, p. 13).

We need to rely on the word of an explorer returning from strange foreign lands because we cannot have exactly *his* reasons for believing what he says. 'We saw it with our own eyes' cannot be our reason for believing what he says. Unless we go there ourselves, we must rely on his telling us what he saw.

With morals, it might seem to be otherwise: we give *reasons* for our moral views and these reasons can be shared and so can be evaluated by each of us.[5] As a crude first approximation, to be refined later, *testifying* that *p* contrasts with *arguing* that *p* insofar as it is the testifier herself who vouches for the truth of *p*; someone who argues that *p* lets the arguments vouch for themselves.[6] So long as we can give arguments for our moral positions, it seems that we do not *ourselves* have to vouch for the truth of those positions. On this view, moral testimony can have a significant place only for beginners who cannot manage to evaluate the arguments. For the rest of us, moral testimony can be useful at most only when we do not have enough time to give careful consideration to all the arguments relevant to a moral position. We might, then, because of time constraints, rely on the moral testimony of those whose character we trust and who we know have been able to give careful consideration to the arguments.

Three additional considerations seem to roll back even further the place for testimony about moral matters.

First, moral matters are important and engage our responsibility in a way that nonmoral matters do not. We are responsible for avoiding errors here in a way that we are not responsible for avoiding them in other areas. Further, moral judgments reflect on our character in ways that other judgments do not. The extra importance attached to moral judgments derives both from their close connection with action and from the importance of the concerns that morality addresses.

---

[5] C. A. J. Coady mentions this as an important consideration against testimony regarding moral matters; see his *Testimony: A Philosophical Study* (New York: Oxford, 1992), p. 71.

[6] This account of testifying is not quite the same as that offered by Coady. He builds speaker competence into his definition of testifying (*op. cit.*, p. 43). Elizabeth Fricker rightly takes him to task for this, since it shifts the interesting epistemological questions to the question of whether someone was in fact testifying that *p*—"Telling and Trusting: Reductionism and Anti-reductionism in the Epistemology of Testimony," *Mind*, civ (1995): 393-41

Second, moral knowledge is supposed to be *practical* knowledge. It is not enough simply to know, for example, that sexism is wrong. One must be able to put that knowledge into practice.[7] But, insofar as one must rely on other people's judgment as to when 'sexism' applies, one's knowledge of the wrongness of sexism will be practically deficient. Similarly, unless one knows not only *that* sexism is wrong but also *why* it is wrong, one will not be able to generalize from the wrongness of sexism to the wrongness of other sorts of discrimination. Because of this, one's knowledge that sexism is wrong will not reverberate throughout one's actions and will not extend itself to other practical spheres.

Third, borrowing moral knowledge appears to be in tension with the ideal of an autonomous moral agent. On some accounts of autonomy, the tension will be straightforward. If we think of autonomy as requiring independence from others, then those who borrow moral knowledge must, by that fact alone, fail to be autonomous.[8] On this picture, being autonomous presupposes what we might call "intellectual individualism."[9] Such accounts of autonomy have been roundly criticized, for it is hard to see why autonomy, conceived as independence, should be something valuable.

A more plausible account of autonomy views autonomy as involving, among other things, the capacity to reflect critically on the principles that one adopts. Through such reflection, one comes to make those principles one's own.[10] But it is easy to see why this account of autonomy, while not as obviously incompatible with borrowing moral knowledge as autonomy-as-independence views, might nonetheless be in tension with it. If we have borrowed our moral knowledge from others and cannot offer reasons in support of it, then how can we reflect critically on that knowledge? Moreoever, if we do not understand the rationale behind those borrowed principles, how can we use our borrowed principles to reflect critically on other principles we accept? How do we know the ways in which those principles connect with others, or the ways in which the considerations that support them align or fail to align with the considerations that we take to support our other principles? And if we cannot bring our borrowed knowledge into play in rationally assessing our other moral commitments, then those commitments will not be given the critical scrutiny they deserve.

---

[7] This point is made by G. E. M. Anscombe, in "Authority in Morals," reprinted in *The Collected Philosophical Papers of G. E. M Anscombe* (Minneapolis Minnesota UP, 1981), pp. 43-50, here pp. 47-48.

[8] For this view, see Wolff; and Thomas Scanlon, "A Theory of Freedom of Expression," *Philosophy and Public Affairs*, I (1972): 204-26. For cogent criticisms of it, see Dworkin, pp. 22-25.

[9] This term is from Sigurdur Kristinsson's Ph D. dissertation, *The Nature and Value of Individual Autonomy* (Cornell University, 1996), p. 129.

[10] Dworkin gestures toward such a requirement (*op. cit.*, p 20).

We began by thinking that since moral views are held for reasons, moral argument should be our primary source for moral knowledge. Testimony can have a place, though a small one, because we do not always have time to examine adequately the reasons in favor of a moral view. Now it looks as though it is incumbent on us, as conscientious agents, to limit our dependency on the moral testimony of others as much as we can. If, in the end, we sometimes still have to rely on testimony, that is a regrettable consequence of our finiteness.

## II. AN EXAMPLE THAT REOPENS THE CASE AGAINST MORAL TESTIMONY

The case against testimony about moral matters appears to be strong. It is easy to construct examples that show, however, that the case against testimony must be less than fully compelling. Here is one such example.

Peter had lived in a fourteen-person cooperative house for two years and was deeply committed to the ideal of cooperative living. His first year there, he had felt uneasy about the cooperative's membership decisions, but the second year he found them intolerable. Membership decisions were made on the basis of quite lengthy interviews, and decisions to accept someone had to be unanimous. Any member could exercise a veto over a potential member for any reason, though in practice such vetoes were rarely necessary, as support for a potential member who might have incurred a veto generally evaporated. Peter came to believe that the threat of veto was making membership decisions arbitrary and unjust. White women and women of color had advocated rejecting three white men on the basis of their perceived sexism, and, in one case, racism. Peter had a settled and serious commitment to the elimination of racism and sexism, but he was not very good at picking out instances of sexism and racism. Thus, his commitment often remained theoretical and he saw fewer occasions as calling for antisexist or antiracist response than he might have seen. Such blindness can sometimes indicate insincerity, but in Peter's case it did not. He genuinely did want to understand the perspective of the women and he wanted to be able to share it, *if* indeed it turned out to be correct. He could pick out egregious instances of sexism and racism, and could sometimes see that "sexist" or "racist" applied to more subtle instances when the reason for their application was explained to him, but he seemed bad at working out how to go on to apply the words to nonegregious new cases. Problems arose for Peter when he could not grasp the reasons why the women were calling someone sexist, when he either could not see, or could not see as evidence, the considerations that the women thought supported viewing the would-be members as sexist. Discussion tended to trail off:

> "No, it wasn't anything he said, it was the way he looked at me when I asked him a question."
>
> "How do you mean?"
>
> "Well, it wasn't that he didn't look at me and he didn't stare at my breasts, it was just a way of looking—a way of not quite being present as he answered me."
>
> "What way? I do not understand what way you mean."
>
> "Well, it was just a feeling I got. A feeling that he didn't think me important."
>
> "But you can't reject someone for a feeling."
>
> "I'm not, I'm rejecting him because he's sexist."

Peter decided to leave the co-op. His reason for leaving was that, as a member, he felt he had to be able to endorse the decisions that the group was making—they had to be decisions that were "his" in the sense that he could identify with them, even if they were the result of compromise and negotiation.

The women in the house wanted to say, and did say: "Look, just trust us, we know about this." But to Peter, such trusting was an abdication of moral responsibility. Even worse, from Peter's point of view, the decisions about which men to accept were arbitrary and discriminated against the men rejected. Nor could such discrimination be justified on the basis of some goal it might achieve; it was not, for instance, a matter of increasing the presence of women and people of color in the co-op. It was, from his point of view, prejudice, pure and simple.

I think that in this case Peter should have been willing to accept the women's testimony that these men were sexist. But for him to have done so would have been to accept and act upon a moral judgment on someone else's say-so. It would have been to borrow moral knowledge.

At this point, it might be objected that the case is misdescribed. Peter was not being asked to rely on other people's moral judgment at all. Instead, he was being asked simply to accept that a purely descriptive term, 'sexist', applied. If we are inclined to think it not *outrageous* that he should have deferred to the women of the house, that is because it is reasonable to suppose that they might be better able than he to apply the descriptive terms 'sexist' and 'sexism'. But, the objection continues, this is no different from our deferring to biologists over when species terms apply, to astronomers over astronomy terms, and so on. It is true that 'sexism' also carries with it connotations of disapproval, but we can divide the meaning of the term into an expressive or condemnatory component and a purely descriptive component. The most popular candidate for the descriptive component would be 'differentiating between the sexes on the basis of their sex and nothing else'.

This objection is mistaken, however. 'Sexism' refers not to just *any* sort of differentiation between the sexes, but to morally unjustifiable differentiation. (It is a failure to grasp this that leads some to say that *any* women-only event has to be sexist since it treats men and women differently simply on the basis of sex.) Thus, to know how to apply the term 'sexism' is to have some moral knowledge.[11]

The objection does point out an important feature of the case, however, and one that shows it to be the simplest, and most readily defended, sort of case. Peter already *had* a commitment to ending sexism. He was not borrowing from others the very principle "sexism is wrong." That much, let us suppose, he had learned *for* himself (though probably not *by* himself). If asked, he would have been able to give reasons in support of it—he might, for example, have said that sexism shows a failure to accord all persons equal respect. Peter did not need to borrow his knowledge of the relevant principle. What he needed to accept on other people's say-so was that this principle applied. We can thus distinguish between taking it on trust that a principle you independently acknowledge applies in a particular case, and taking it on trust that a principle is correct. In between these two, we can also identify taking it on trust that the principles you recognize on independent grounds are to be ranked in a given way.[12] (Such rankings are often at issue in setting political agendas.) We shall need to consider whether there is reason to be more or less suspicious of borrowing moral knowledge depending on what kind of knowledge it is that you are borrowing.

---

[11] Sometimes there is a clear answer to the question of whether something is moral knowledge or the sort of nonmoral knowledge that feeds into moral knowledge. (For example, average daily caloric intake of the poor may be the sort of factor on which the justice of a society supervenes, but it is clearly nonmoral.) But I do not think this distinction is always easy to draw. For example, should we say that the fact that someone will be upset by a course of action is a purely nonevaluative fact, though one which may contribute to determining whether a course of action is right or wrong? Or should we say it is itself an evaluative consideration? There is reason for thinking it is evaluative. Being upset contrasts with what is colloquially called "having a hissy fit," since the later implies that the distress is prima facie unjustified. 'Upset' may thus have connotations of prima facie justified (or at any rate not prima facie unjustified) distress. It is worth noting that, if you want to maintain a sharp separation between moral and nonmoral epistemology and allow testimony to the latter but not the former, then you had better be able to draw clear lines between the moral and the nonmoral. A view like my own, that postulates no such sharp dichotomy at the level of epistemology, need not worry about being able to draw this distinction in every case.

While I think that Peter was mistaken to refuse to accept the women's testimony regarding whether someone was sexist, his refusal nonetheless displayed admirable traits of character and admirable strength of character. Peter's action was not motivated by a desire to control the outcome of collective decisions. He made the decision to leave quietly and it was never intended as a threat to force the group to decide in the way he thought correct. The esteem in which Peter and the women held each other made it possible to sustain friendships despite the disagreement and despite Peter's leaving. It was quite probable, too, that others shared Peter's misgivings, but lacked his conscientiousness about moral matters.

Peter's stance shows us the prima facie plausibility of the three considerations against accepting moral testimony: he shows a keen and admirable awareness of the importance of moral issues. Moreover, his desire to have, *and to understand*, morally adequate reasons for acting makes us fairly sure, that, should Peter fall into morally bad circumstances, he will not be following the leader into error. This sort of independence—the independence of those who want to keep their moral beliefs in step with what they take to be cogent reasons—gives us a sense of the notion of autonomy which shows why autonomy is thought to be something valuable. Finally, had Peter accepted the women's testimony, he would still not have been any the wiser about how to go on to apply 'sexism' to new, nonegregious, cases.

Overall, though, Peter's case does more to undermine the arguments against testimony than to support them. In particular, it invites us to reexamine the initial assumption that led us to believe that testimony must, for the morally mature, always be an inferior way of learning about morality relative to learning about it through argument. We are not supposed to need testimony, except when time is short, because someone who accepts a moral position can always make the grounds of her beliefs accessible to us, and we can come to share her reasons or not, depending on whether we find them persuasive.

---

[12] This tripartite distinction does not map neatly onto Gibbard's distinction between fundamental and contextual authority (*op. cit.*, pp. 174-81). Contextual authority is authority granted to another on the presupposition of shared norms Gibbard's examples, however, suggest that he may have in mind cases in which the borrower is ignorant about some nonmoral facts of the situation and so lets someone else's moral reasoning stand "proxy" (*op. cit*, p. 174) for her own. A moral agent claims fundamental authority when she claims the superior trustworthiness of her own norms. But it seems that one can accord another *some* fundamental authority without accepting their norms on the basis of her say-so, as when one takes the fact that someone believes a norm as *a* reason for accepting it, but as a reason that is, in the circumstances, outweighed (*op. cit.*, p. 180).

Reflection on Peter's case suggests two compatible responses to this argument. We might say that at least sometimes, the capacities involved in acquiring moral knowledge are indeed similar to perceptual capacities. There is a moral analogue of blindness, which is not to be assimilated to being too foolish to follow a moral argument. Those who lack the relevant perceptual skills in a given moral domain must rely on those who have sharper moral vision.

Alternatively, we might point out that the contrast between testifying and arguing is not hard and fast. Expert witnesses both testify and argue. They offer arguments when presenting the considerations that lead them to adopt their views, but they testify, too, since the nonexpert to whom the testimony is directed is not in a position to assess fully the cogency of the reasons the experts offer in favor of their views. The nonexpert may not know exactly why a certain response on a given test indicates with such-and-such probability that the blood sample comes from the defendant, but will, having established the expert's credibility, take her word that it does. Offering arguments for a position thus need not entirely eliminate the reliance on the word of the testifier which is characteristic of offering testimony. Expert testimony shows the context sensitivity of the notion of testimony—directed at fellow experts, the arguments that for the lay person are also testimony would be just arguments.

I think that both responses have a place and are not as problematic as they might first appear. In defense of the claim that moral knowledge sometimes requires skills analogous to perception, we need only think of those occasions in which seeing that some moral value is at stake requires an emotional sensitivity to the particulars of a situation. For example, seeing that a situation calls for kindness may require an empathetic response involving imagination of what it must be like to be in someone else's shoes. But we might not all be equally capable of such empathetic leaps—some of us may need to rely on those who are.

Sometimes, those from whom we accept moral testimony stand to us, regarding certain issues, as moral experts. But the very idea of moral experts might seem deeply problematic. It seems profoundly inegalitarian to claim that some are moral experts, since it suggests that others are not.[13] While the higher branches of mathematics might be beyond some of us regardless of our best attempts, and available to others only if they make certain attempts, it is often thought that each of us is equally capable of being a good moral agent. And there can be no experts where capacity is equal.[14] Moreover, an induction on the past history of those who have claimed to be moral experts and with what results might make us skeptical of the very idea of moral experts. Many who have claimed that role have used it to advance self-serving and morally dubious views.

These concerns about moral experts need not be taken as concerns about the very idea of moral expertise. They are for the most part generated by worries about establishing a socially sanctioned role of moral expert. Having a socially sanctioned role of moral expert is going to raise issues of credentialing—Who are the moral experts and who gets to decide?—and issues of what authority those in the socially sanctioned role of moral expert are to have over policy formation, and over the rest of us who may disagree with their judgments. These issues, however, do not arise when considering the question of what sorts of circumstances, if any, are ones in which an agent should decide to rely on someone else's moral judgment.

We can further reduce the sting of acknowledging that there can be expert testimony on moral matters by distinguishing between *experts* and *expertise*,[15] and between *global* and *local* claims to expertise. While we think that there can be few experts on any subject—if your knowledge is not rare and hard to come by, you would not merit the title "expert"—expertise can be widely shared. Normal adults have considerable expertise at a wide variety of things. To borrow a list from Arthur Caplan, "how to drive a car, mail a letter, exhibit good manners, use a post office, play games" (*op. cit.*, pp. 30-31), and, we might add, considerable expertise in ethical matters. Thus, in saying that someone can sometimes stand to another much as expert witness to lay person, we are not committed to saying that the person who currently does not have the expertise could not acquire it, though in some cases the kinds of character change and changes in experience necessary in order for them to acquire it may be sufficiently radical that, for all practical purposes, they are unable to do so. More important, we are not committed to saying that one person stands to another as expert to lay person in *all* areas of value. If that were indeed the case, it would be tempting to say that the one forced to rely on the other would lack moral maturity, so we would be talking not about a case in which the morally mature agent needed to borrow moral knowledge but about a case in which the agent was still a beginner.

---

[13] See Arthur Caplan, "Moral Experts and Moral Expertise: Does Either Exist?" reprinted in his *If I Were a Rich Man Could I Buy a Pancreas?* (Bloomington: Indiana UP, 1992), pp. 18-39.

[14] This thought lies behind the strong suspicion that if, for example, Lawrence Kohlberg's scale of moral development consistently scores women, the working class, and members of other cultures at a lower level of moral development than educated middle-class white males, this must indicate some bias in the scale. The suspicion of bias is not removed by the fact that when education and type of employment are controlled for, men and women score the same. In contrast, we would not infer bias in a scale that aimed to measure mathematical skill if it led to similar results. We would simply say that those with certain educational backgrounds and occupations were, most likely, better mathematicians than those who lacked these backgrounds. While it might be plausible that having a certain background exposes you to certain *types* of moral problems, however, everyone, we think, has exposure to moral problems of some type or other, and thus everyone will have the sort of experience necessary to develop their potential as moral agents, though this development may take different forms.

[15] For this distinction, see Caplan.

Although lowering the cost of the analogy between expert testimony and testimony regarding morality, none of what has been said so far gives us any reason to suppose that moral expertise might be unevenly distributed with respect to certain domains.[16] But evidence for this claim is not hard to find. Examples like that of Peter provide us with the sort of evidence we need. Nor does positing such differences in likely familiarity and knowledge commit us to any epistemologically radical claims. Baier (*op. cit.*, p. 247) and others argue that one's moral experience shapes the kinds of questions one will find salient: had moral theorists begun from the experience of women, they would have paid more attention to trust, since trust is of vital importance in relations among unequals.

Political engagement in liberation movements can also be expected to lead to greater insight and understanding about the values at stake than is likely to be available to someone who lacked that sort of experience. One way to find out about the value of respect is to ask those who have always been respected; a better way is to ask those who have struggled to win respect and tried to live without it. Saying that such political experience can contribute to moral knowledge commits us to nothing stronger than the view that one's knowledge about moral matters, like one's knowledge about other matters, can deepen with experience. In particular, the suggestion that, for example, women actively engaged in feminist movement, might have acquired some expertise about the kind of disvalue sexism is, and the often subtle forms it can take, does not require us to defend a notion of epistemic privilege of the kind defended by standpoint theorists.[17] It is enough if, through social location or choice, individuals come to have richer experiences with certain types of moral problems. We need not follow standpoint theorists and suppose in addition that the marginalized are always, or even generally, in the best position to understand the truth about social relations.[18]

---

[16] In fact, this would have to be concerning certain areas only, or we would run afoul of the idea that, globally, there will not be much difference in moral skill between people in distinct groups.

If we allow that capacities akin to perceptual capacities are sometimes involved in generating moral knowledge—especially the knowledge of when values are at stake in particular situations—and if we admit that there can be local differences in moral expertise, differences explained by different experiences, then we have come some way toward showing that there can be room for moral testimony. Sometimes, we do need to hear, and to believe, reports of explorers returning from realms of value that are less familiar to us than to them. Nonetheless, one might continue to be uneasy about moral testimony: What of the objections that borrowing moral knowledge underestimates the importance of moral matters, overlooks their practicality, and conflicts with autonomy? The first of these objections is best dealt with now, while the second and third need to wait until we have a clearer understanding of when an agent is justified in relying on moral testimony.

We need not suppose that we have to meet the first objection on exactly its own ground. There is a good case to be made for just denying that there is any difference between moral and other knowledge with respect to the way it engages our responsibility: we are every bit as accountable for believing the racist science of *The Bell Curve*[19] as we are for having false moral beliefs. But the strongest case against this position can be made by affirming its premise. If it is so very bad to make a moral mistake, then it would take astonishing arrogance to suppose that this supports a do-it-yourself approach, especially once we think about how expertise can vary with experience. Of course, the worry behind this objection might be that people will too readily rest content with borrowing the moral beliefs of others without any grounds for supposing them likely to be correct. And that, the objection goes, would be a failure to take moral matters as seriously as they should be taken. This assumes, however, that the borrower of moral knowledge has few responsibilities to meet before she can count as borrowing wisely. In what follows, I shall argue that the standards for wise borrowing are stringent indeed and so offer no comfort to the morally indolent.

---

[17] For this use of 'feminist movement' rather than 'the feminist movement', see bell hooks, *Feminist Theory from Margin to Center* (Boston: South End, 1984).

[18] For a discussion of standpoint theory, see Nancy Hartsock, "The Feminist Standpoint: Developing the Ground for a Specifically Feminist Historical Materialism," in Sandra Harding and Merrill B. Hintikka, eds., *Discovering Reality. Feminist Perspectives on Epistemology, Metaphysics, Methodology, and Philosophy of Science* (Dordrecht: Reidel, 1983), pp. 283-310. It would be a mistake to think that political engagement must count against objectivity. Even if such engagement means that one will approach questions with presuppositions, provided that one's presuppositions are true (or approximately true), one can count as being objective. Note that unless this were the case, the theory dependence of scientific method would count against the objectivity of science. For this notion of objectivity, see Richard Boyd, "How to Be Moral Realist," in G. Sayre-McCord, ed., *Essays on Moral Realism* (Ithaca: Cornell, 1988), pp. 181-228; and Louise Antony, "Quine as Feminist: The Radical Import of Naturalized Epistemology," in Antony and Charlotte Witt, eds., *A Mind of One's Own* (Boulder: Westview, 1993), pp. 185-225.

[19] Richard J. Herrnstein and Charles Murray, *The Bell Curve. Intelligence and Class Structure in American Life* (New York: Free Press, 1994).

## III WISE TRUSTING

I shall argue that the problem of how to be a wise receiver of moral testimony is a special case of the problem of how to be a wise receiver of testimony more generally, and this in turn is a special case of the problem of how to be wise in one's trust. Because of this, theories of trust, and in particular accounts of trust's justification conditions, can offer insight into some of the epistemological issues raised by testimony. But first it is necessary to show that relying on testimony does, in the typical case, require trust in the testifier.

It might seem simply obvious that testimony involves trust: where there is testimony there is trust, or so says common sense. The language in which we describe cases of testimony supports this claim, for we speak of "taking it on trust," and of "trusting testimony." While I shall argue that common sense is basically right here, it does take some additional argument to vindicate it—unless, that is, we have a minimalist account of trust.

So as not to mislead the reader into thinking the connection between trust and testimony follows just as a matter of course, I shall temporarily abstain from using phrases like 'trusting testimony' or 'taking it on trust', and use instead the phrase 'accepting testimony'. Someone accepts the testimony of another if she forms the belief that *p* on the basis of the other's saying that *p* is the case. We can *use* testimony that we do not accept: for example, the movie critic whose rave tells us that we would hate the movie, or the suspect whose answer signals that she has something to hide. The thesis that needs to be defended can be put like this: in the typical case, accepting testimony requires trust in the testifier and, as often as not, in others as well.

How plausible one finds the claim that there is a connection between trust and accepting testimony depends, not surprisingly, on the account of trust one accepts. If you think that trust is just a matter of relying on others, then it is obvious that accepting testimony has got something to do with trust. Call accounts of this kind *risk-assessment accounts*. Diego Gambetta[20] offers us an example of a risk-assessment account:

> . .trust (or, symmetrically, distrust) is a certain level of subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both *before* he can monitor such action (or independently of his capacity ever to be able to monitor it) *and* in a context in which it affects *his own* action (*ibid.*, p. 217).

On a risk-assessment account, it follows straightforwardly that accepting testimony requires trust: I would not accept your testimony unless I assigned a sufficiently high degree of probability to your telling me the truth in these circumstances.

Risk-assessment accounts do not distinguish among the reasons one might assign a sufficiently high degree of probability to an agent's performing an action. One might think an agent will perform that action out of fear or stupidity, because it coincides with her own self-interest, or because she has, and wishes to display, goodwill toward those who are counting on her.

Will-based accounts, in contrast, find trust only where there is reliance on the goodwill of another. We are not to take 'goodwill' as synonymous with 'personal friendly feeling', or will-based accounts would be obviously too narrow. Sometimes those we trust have the relevant goodwill just in virtue of being morally decent, or honest, or caring about fulfilling their duty. Since will-based accounts are more restrictive than risk-assessment accounts, if I can show that even on such accounts trust is needed, then I shall have shown that *whatever* account of trust is ultimately endorsed, accepting testimony requires trust. Consider the following will-based account of trust according to which trust is:

> (a) an attitude of optimism that the goodwill and competence of another will extend to cover the domain of your interaction, together with
>
> (b) the confident expectation that the one trusted will be directly and favorably moved by the thought that you are counting on her.[21]

There are clearly some nonstandard cases in which I can accept testimony without having an attitude of optimism that the goodwill of another will extend to cover the domain of our interaction. I may, for example, have evidence that you would not dare lie because I have a threat advantage, and know that you are someone who is susceptible to threats. In a different scenario, I might know that it will somehow suit your devious purposes to tell me the truth, here and now (though in general you would not). Usually, however, we are in neither position: we have no threat advantage sufficient to guarantee performance without also relying on goodwill, and we do not have the kind of information that would let us know when those who are generally untrustworthy are in fact telling the truth. The false witness, like the lying promise giver, is threatened with the possibility of never again having her testimony believed.[22] But this threat is not sufficient to guarantee performance in such a way as to eliminate the need for reliance on goodwill, and to be effective at all it requires both that we be able to tell when we have been deceived and that we will again encounter the false witness.[23] These conditions are frequently not met.

---

[20] "Can We Trust Trust?" in Gambetta, ed., *Trust: Making and Breaking Cooperative Relations* (New York: Blackwell, 1988), pp. 213-38.

[21] For a defense of this account, see my "Trust as an Affective Attitude," *Ethics*, CVII (1996): 4-25. Baier defends a will-based account (*op. cit.*).

What about the need for optimism about the competence of another? In accepting testimony, we typically lack the kind of proof about the competence of another that would securely ground our optimism in her competence. Indeed, certain kinds of proof are ruled out. If testimony is to be at all useful, I cannot myself go and check up on what you are saying. Sometimes, my not doing so is merely a matter of convenience: I accept testimony precisely because it would be too bothersome to learn for myself what I can learn from you. But where testimony has the potential to have rich epistemic yield, I am not in a position to do the kind of checking up on your judgment which would be required. I may be ignorant about, for example, some mathematical theorem, and lack the expertise to find out the truth of the matter on my own. Even if I attempted to check up on your word by trying to prove it on my own, I, novice that I am, shall often have fewer grounds for confidence in my own judgment than I have for confidence in yours.[24]

At this point, it might be objected that I have failed to distinguish two kinds of checking up: I may not be able to check up on the *content* of what you say, but I may be able to check up on *you*.[25] We often do call on other people's assessment of our informant's character and expertise before accepting their testimony. But this move does not eliminate trust; rather, it grounds my trust in you in my trust in others. It may disperse trust, but it does not make it redundant.

---

[22] For this penalty in the case of the false promisor, see David Hume, *A Treatise of Human Nature*, P. H. Nidditch, ed. (New York: Oxford, 1978, 2d ed.), p. 522.

[23] For this response to the objection that trust is not needed in the scientific context because game-theoretic considerations alone will be sufficient to guarantee honesty, see Hardwig, p. 704.

[24] Hardwig makes this point (*op. cit.*, pp. 698-99).

[25] These are distinguished by Dworkin (*op. cit.*, p. 56).

To show that it is true on the account of trust in question that accepting testimony typically requires trust in the testifier, we also need to show that such acceptance will require the confident expectation that the one trusted will be directly and favorably moved by the thought that one who trusts is counting on them. It is easy to show that this is also required: conversational exchanges do not come labeled, conveniently announcing themselves as the kinds of exchanges that they are. Before I can accept what you say in response to my request or need for information, I must first assume that what you are doing is *in fact* responding to my request or need—that you are not, for example, teasing me by saying something silly since you think a silly question deserves a silly answer.

The common-sense assumption that accepting testimony involves trust is borne out, if not for every possible case, then for the standard cases and these form the vast majority of cases of accepting testimony.

If this is right, then the question of what it is to be a wise and responsible recipient of moral testimony just will *be* the question of what it takes to trust wisely in these domains. And that is a nice result, because we know something about the justification conditions of trust.

In *Testimony*, C. A. J. Coady tries to argue that our default stance should be one of trust in testimony: we require some evidence to move away from this stance and none to adopt it. Viewed from the perspective of accounts of when trust is justified, this claim invites suspicion. It is simply impossible to say what our default stance with respect to trust in general should be. For this reason, we should be surprised indeed to find that one could defend a default stance of trust with respect to testimony, given that testimony can cover a wide range of areas, be given by a wide range of testifiers, and occur in different social climates and different institutional settings.[28]

---

[28] Fricker notes the context sensitivity of appropriate trust in testimony (*op. cit.*). She accuses Coady of an inconsistency in saying that our default stance should be one of trust and at the same time allowing that we accept testimony only when signs of untrustworthiness are absent (*op. cit.*, p. 405). But this is still to have a default stance of trust. So long as I do not need positive evidence of trustworthiness before I trust, then my default stance is one of trust. In what follows, I argue that sometimes our default stance can be distrust; thus my position may be more skeptical than Fricker's.

The reason there are no useful generalizations to be made concerning whether trust, distrust, or neutrality is the appropriate default stance is that the appropriate default stance is sensitive to climate, to domain, and, linked to domain, to consequences. Further, for the individual truster, the appropriate default stance is linked to the reasonable assessment of the tendencies of her own trusting and distrusting.[27]

In climates in which the there is strong motive to be untrustworthy, it would require more evidence to believe that a person was trustworthy than it would take to believe that a person is trustworthy in climates where there is little incentive to untrustworthiness. Thus in unfavorable climates, our default stance should be distrust, and it should take considerable evidence to move from that position to trusting; in more favorable climates, however, a default stance of trust can sometimes be warranted.

Domain also affects which default stance is appropriate: if I am to trust you with a secret of mine, then I shall want to have quite a bit of evidence about your character. If, though, I am to trust you not to attack me in the street, I may need no particular evidence about your character at all. This might seem to be a counterintuitive result, since surely it is worse to be attacked in the street than it is to be embarrassed by a confidence indiscreetly betrayed. Domain is generally more important than the consequences of misplaced trust, however, because domain signals likelihood of performance. We are all aware of the lively attractions of gossip and the strength of character required to resist them. In contrast, it is not hard to refrain from injuring a stranger on the street; that just takes basic decency, a trait that we can assume is widely shared, unless the climate is sufficiently bad. Once we hold domain fixed, consequences become of the first importance: of course, I am going to need more evidence of your trustworthiness before I tell you a secret that, if spread abroad, would be very damaging to me, than before I tell you a secret whose disclosure would be merely embarrassing.[28]

---

[27] These conditions for justified trusting are further elaborated and defended in my "Trust as an Affective Attitude," especially at pp. 20-25.

[28] Hardwig wonders about the extent to which the truster of scientific testimony needs to have good evidence for her belief that the inquirer is trustworthy (*op. cit.*, p. 699). It seems to me that his account would be more plausible if he were to take differences in domain into account. Scientific inquiry has an historically poor track record concerning politically sensitive topics, such as race and gender. Whatever we are to say about the kinds of evidence needed before accepting the research of other scientists in general, it seems we need extra caution when it comes to research that can validate social relations of dominance and subordination

While climate, domain, and consequences are variables determining which default stance is justified that extend across agents, the fourth variable determining the appropriate default stance is agent specific. Some agents have reason to be distrustful of their tendencies toward trust in certain domains. Others have reason to be distrustful of their tendencies toward distrust. Here we might think of, for example, racist tendencies to distrust young African-American men. Such agents should be suspicious of their own default stances, and that, in turn, should lead them to be more willing to abandon them than those who can trust their trust and distrust.

If climate, domain, consequences, and metastances of trust or distrust are the important variables determining how much evidence is warranted before trusting, we can see why there can be no answer as to the appropriate stance to take toward testimony in general. Most people can be trusted to give an honest and informed answer to certain questions, including the mundane sorts of questions we often rely on strangers to answer for us, such as what time it is, how we get to the bus stop, and so on.[29] But when it comes to matters where there are reasons for not being straightforward—for example, social-science surveys of sexual activities—or to matters where competence cannot be assumed, we should not adopt a default stance of trust in testimony.

This general discussion has implications for what we should say about how much and what sort of evidence we should demand before trusting someone's moral testimony. Moral knowledge regarding the sorts of matters for which we might need moral testimony cannot be easy to come by. If it were easy to come by, then we would expect it to be in the commonly shared stock of moral expertise. In addition, our moral thinking is vulnerable to distortion due to, for example, wish fulfillment and self-interest. Having such knowledge requires a good character, as well as the right sorts of experiences. Since the best way to convince others to go along with your interests is to convince them that morality requires them to do so, we can expect tendencies toward untrustworthiness—perhaps deliberate, perhaps as the result of self-deception.

If this is right, then the appropriate default stance toward testifiers about morality is one of distrust. A default stance of distrust gives a presumption against accepting moral testimony, where this presumption can take the form of either disbelieving such testimony, or withholding judgment. What distinguishes a default stance of distrust from a stance of neutrality is that a substantial weight of evidence is required to rebut the presumption; with neutrality, there is no presumption that stands in need of rebutting.

---

[29] But even this list is climate sensitive. I am told that in Egypt it is thought disrespectful not even to attempt to offer directions to those who ask for them. Rather than be disrespectful, an informant will send you in completely the wrong direction.

In discussing Peter's case, I conceded that it was the simplest sort of case, since it involved relying on someone else's judgment that a principle he already accepted applied. If your source is wrong about the relative importance of a good, or wrong about which principles or ends to adopt, *and* you take her word for it, your error is going to surface not just in one place but in many; if, however, you rely on someone's judgment that a principle you accept on independent grounds is at stake in a given situation and she is wrong, at least the error you inherit is contained. The intuition that these kinds of cases are importantly different is shown to be just a special instance of the way in which consequences interact with domain to affect the justifiedness of trust.

But acknowledging that the appropriate default stance toward testifiers on moral matters is distrust does not mean that we can never be licensed in moving from distrust to a position of trust. Before we do so, we shall want to have good evidence about the person's character, about possible hidden agendas, and about whether she has the sort of experiences that contribute to the kind of competence we are counting on her to have. We would also want to know that our witnesses have appropriate epistemic self-assessment,[30] and are not given to asserting with confidence claims that they are in a position to assert only tentatively if at all. Perhaps the most trustworthy testifiers about moral matters are those who are least inclined to offer such testimony.

Returning to Peter's case, we can see, I think, that had Peter accepted the women's testimony, his trust would have been justified. Having lived in the same household, Peter had available enough of the right kind of information to justify moving from a default position of distrust to one of trust. Both sides thought well of each other's character.

In addition, Peter had reasons to distrust his own distrust. It might be objected that Peter could not know this in advance of knowing whether the women were in fact correct. But if Peter could have known whether the women were correct, then there would have been no need to trust them in the first place. This objection forces us to consider whether the notion of appropriate metastance can pull enough independent weight to enter into our assessment of when trust is justified, or whether our assessment of appropriate metastance must always be consequent on our assessment of appropriate first-order trust.

---

[30] For the notion of epistemic self-assessment, see Hardwig, p. 700.

I think, though, that the example itself shows why this objection is mistaken. Peter had plenty of evidence to make him suspicious of his distrust. His past track record reveals a tendency to find fewer occasions calling for antiracist and antisexist response than others. Moreover, given what Peter knew, or should have known, about how experience affects competence in these matters, he should have been aware that others might be better placed to judge than himself.

In the example, I assumed the women were in broad agreement about whether there was evidence of attitudinal sexism. Had there not been such agreement, though, it is still possible that Peter would have been justified in trusting the women who found evidence of sexism. Agreement is by no means a sure indicator of truth. Sometimes it counts against the trustworthiness of testifiers, signaling that they are in cahoots, rather than independently correct.[31] And sometimes, especially in political contexts, it counts against the trustworthiness of testifiers by signaling that some voices have yet to be heard. In fact, given the complex interests that are so crudely captured within identity politics, unanimity among the members of an oppressed group is generally counterevidence for trustworthiness. One of the chief responsibilities of wise trusters will thus be determining whom to trust when there is disagreement.

The evidential standards that must be met before trust in someone else's moral testimony is justified are high. This means that the morally indolent can take no particular comfort from a defense of the possibility of borrowing other people's moral knowledge. There is still work to be done and responsibilities to be met before we trust testifiers on moral matters wisely. But we can see now how to respond to one of our outstanding worries, namely, the concern that accepting moral testimony conflicts with autonomy. We must subject judgments about whom we trust on these matters to rational scrutiny, and we can be called on to revise our assessments of who is trustworthy. Accepting moral testimony is shown not to be in conflict with the value of autonomy, properly understood.

We might still wonder, though, whether, had Peter relented and trusted the testimony of the women, there would have been something inferior about his epistemic status. There are two questions here: Could Peter have actually come to know through trusting? Is his knowledge deficient as practical knowledge?

---

[31] See Coady, p. 192.

While I think that wise trusters can have knowledge on the basis of testimony—provided, that is, they are lucky enough to trust trustworthy knowers—I am going to leave that issue to one side so as to take up the issue of practicality.[32] Even if Peter had been willing to borrow the women's moral knowledge, at the end of the day he would have been as much in the dark as to how to go about applying 'sexism' to new nonegregious cases as he ever was. Of course, we can hope that with discussion and practice he will increase his sensitivity. There is, though, no guarantee that this will happen. Borrowing may be the best that Peter can ever do.

So does this not show that there is something regrettable about the need to borrow moral knowledge? At most, it shows there are some cases in which it is regrettable. Consider now a case of a different sort, concerning ranking principles or ends. The Chinese author Luxun once quipped that the Chinese do not understand freedom, they only understand cooking oil. Let us start from this quip and construct a case. It is China in the early twentieth century and two long-time friends are discussing whether to join a political movement for democratic freedom. One has returned from study in Europe and has himself joined and is trying to persuade the other to do so. His friend is reluctant: "It is dangerous," he says. "Don't we have responsibilities to maintain certain traditions? And isn't the very idea of such change altogether too European?" The one who has returned from Europe tries to convince his friend that he does not know what freedom is really like and that, if he did, he, too, would consider the risk worth it and the tradition stifling. "You haven't," he says, "tasted freedom, and if you had, you would choose as I do." They talk, we might suppose, into the small hours of many nights. Finally, the other also decides to join the political movement, because of his trust in his friend. But that choice does not leave everything as it was before. Having made the choice, he is thrown into different circumstances and comes to see things differently. Through pursuing democratic freedom, he comes to understand more fully its value. He now knows first-hand what he initially knew only second-hand, and his belief, originally grounded in testimony, comes to have rich alternative sources of support.

---

[32] Briefly, the argument for this claim is as follows: in accepting your testimony, I am attempting to borrow your knowledge, that is, to claim the knowledge you have as my own on the basis of your having shared it with me. But I cannot borrow from you what you do not have. Thus, before my accepting your testimony can be knowledge conferring, you yourself must have knowledge regarding the question at issue. If you are wrong, and I take your word for it, I obviously do not know it; likewise, if you are right, but only accidentally right, my being right is no more securely tethered to the truth of the matter than your being right is.

But while your knowing that *p* is a necessary condition for my knowing that *p* on the basis of your say-so, it is not a sufficient condition. Even if you know that *p*, you might not be trustworthy with respect to this exchange of information. For example, you might be telling me that *p* because you think that is a good way to have me come to believe that not-*p*. This would be true, if, for example, you falsely believed that I was acutely suspicious of you and so would take your word to be counterevidence for the truth of *p*. To the extent that my acceptance of testimony rests on trust (and not, for example, on having access to the roundabout reason you happened to tell me the truth), you must be *trustworthy* in the relevant domain before I can know that *p* on the basis of your telling me that *p* If you are not trustworthy, then you could just as well be telling me a falsehood, and I would be in no position to discern this; thus my connection to the truth of *p*, arrived at through your say-so, would be too tenuous to count as giving me knowledge that *p*, even if you yourself knew that *p*

None of this shows that those who trust wisely, and are fortunate enough to trust trustworthy knowers, can thereby have knowledge But, as Coady has persuasively argued, our knowledge is so testimony saturated that, if the wise truster cannot know, then there is little indeed that anyone can know.

It might be objected that the newcomer to the political movement did not need to trust his friend at all and that he was not in fact deferring to his friend's moral expertise—he was just trying it out and there is no deference involved in trying something out. But this reconstruction of the example overlooks two important features of it. First, the newcomer initially had objections to joining the movement, objections that had to be overcome before he could see himself as justified in even merely "trying" it out. Second, without at least some trust in the friend, it would not have been responsible for the newcomer to put himself in circumstances which would predictably lead to changes in his experiences and values, changes which from his current perspective look morally suspicious. It is no comfort to know that once we make a choice, we shall come to have experiences that confirm that choice as the right choice, unless we also believe that such confirmation will be truth conducive. Joining a cult might lead us to have experiences that will confirm our decision—but that we shall have such experiences is no reason to join; rather, it is reason to steer well clear, least our judgment come to be distorted by our membership. The newcomer thus had to trust the friend that membership in the political group would generate experiences that were truth conducive rather than error confirming.

At this point, it might be thought that my defense of testimony remains grudging. It is better when knowledge initially gained through testimony is fully integrated into a new evaluative outlook and the person comes to know first-hand what she originally knew only second-hand. Someone might accept this and yet still maintain that we have an obligation to cultivate all our moral capacities and so use testimony concerning moral matters only as a last resort when there is no other avenue open to us to secure the knowledge we need. Is it not part of living a morally serious life that we keep pushing at the boundaries of that which we must take on trust until the place for testimony shrinks to near vanishing?

The suggestion that we have an obligation to keep pushing back the place for testimony rests on some empirical assumptions, however, that we should hold doubtful: in particular, it assumes that we can develop our moral capacities to an equal extent. Given the world of value is complex and the capacities needed to navigate in it various, there is reason to suppose that we cannot develop all morally relevant sensitivities to an equal extent. And we may do worse by trying than by practicing wise trusting.

These remarks suggest an interesting answer to the question of the unity of the virtues. I appear to be assuming that you can have one virtue without having another. Owen Flanagan[33] argues that psychological research shows the capacities that support the particular virtues do not come in a unified package; thus one can be courageous without being kind, and we should abandon the thesis of the unity of virtue. If one feels the pull of these empirical considerations, yet also feels the pull of a Socratic a priori argument for the unity of the virtues, then hope for reconciliation lies in the notion of wise trusting. Recall the argument for the unity of the virtues goes as follows: virtues are states that issue in only right action. Justice and kindness can *appear* to conflict: if one's kindness is not to issue in wrong action, one must also possess the virtue of justice (kindness that leads to injustice is no real kindness). If we think of a virtue as a sensitivity to considerations of a certain kind, then given that there are no limits on the combinations of such considerations that can be present in a choice situation, such apparent conflict between the virtues is always a possibility.[34] Thus, if the virtues are to result in only right actions, one cannot have one without having them all.

The notion of wise trusting lets us accept the empirical claims typically taken to support the disunity of the virtues, while retaining a kind of unity to the virtues. While the virtuous person might not herself have all the capacities necessary for her virtue to lead to only right action, she will have appropriate self-assessment and appreciate the limits of her capacities. She will know when she needs the moral help of others. Moreover, she will be wise in whom to trust. Her capacities, supplemented by the capacity for wise trust, will issue in only right action.

---

[33] *Varieties of Moral Personality: Ethics and Psychological Realism* (Cambridge: Harvard, 1991).

[34] For this formulation of the Socratic argument, see John McDowell, "Virtue and Reason," *The Monist*, LXII (1979): 331-50, especially pp. 331-33

If these reflections are along the right lines, then we can see another sense in which the virtuous person's virtue is essentially social: it is not just that the social sphere provides much of the domain for the exercise of virtue, nor just that it provides the conditions for sustaining it; rather, the knowledge that the virtuous person has is knowledge made available to her in part through epistemic dependence on others.

KAREN JONES

Cornell University

# RESTITUTION AND REVENGE*

Why punish wrongdoers? I have never found standard retributive accounts of punishment to be able to provide a satisfactory answer. While the intuition that those who do evil deserve to be punished is not incoherent, it just seems to supply insufficient reason to harm wrongdoers.[1] Retribution does not seem to do much good, if any, for either the direct victim, society at large, or the criminal. I suggest that we instead try to answer the original question by understanding punishment as a debt the criminal pays to his victim(s) as compensation. This can obviously explain what good punishment does the victim of the crime. The payment of a debt can also benefit the criminal insomuch as it is a form of atonement. It can be so considered, for the payment provides the criminal an opportunity to make amends for his misconduct and demonstrate his remorse. It can also reduce the pangs of his conscience. The appeal of the atonement/debt approach, however, is not only that it can remove the guilt of the convict in his own eyes, but that it can create in the mind of the public the belief that he deserves to be forgiven or, at least, granted a fresh start with a clean slate since he has made good on his debt. Since this form of punishment can benefit the guilt-stricken wrongdoer, it becomes understandable why it might be said that the criminal has a right to be punished and would even choose on some occasions to exercise it.

An additional advantage of this atonement/debt theory is that it actually provides a place for vindictiveness in our judicial system. The debt the criminal owes his victim can be paid when the latter

---

* I would like to thank Frances Myrna Kamm for a number of helpful conversations on the topic of punishment.

[1] I am not the only one puzzled about how the retributivist idea by itself can provide sufficient grounds for harming criminals. See Warren Quinn, "The Right to Threaten and the Right to Punish," in A. John Simmons, Marshall Cohen, Joshua Cohen, and Charles R. Beitz, eds., *Punishment: A Philosophy and Public Affairs Reader* (Princeton. University Press, 1995), pp. 47–93, here p. 54; David Ross, *The Right and the Good* (New York: Oxford, 1930), pp. 57-60; S.I. Benn and R.S. Peters, "The Utilitarian Case for Deterrence," in David M. Adams, ed., *Philosophical Problems in the Law* (Belmont: Wadsworth, 1996), pp. 416–20, here pp. 416, 419-20; and Jean Hampton, "The Moral Education Theory of Punishment," John Rawls "The Practice of Punishment," Margaret Holmgren, "Punishment as Restitution: The Rights of the Community," David Lyons, "Punishment as Retribution"—all in Michael J. Gorr and Sterling Harwood, eds., *Crime and Punishment. Philosophical Explorations* (Boston: Jones and Bartlett, 1995), pp 356–73, here p 372; 336–42, here p. 338; 402–17, here p. 402; 317–21, here p. 317— respectively.

0022-362X/99/9602/79–94

takes his legal revenge upon the former. The vindictive pleasures accompanying legal punishment can either increase the value of any other form of payment received, or, where the criminal is destitute, can actually take the place of receiving financial compensation. When a debt has been fairly determined and properly paid, both the victim and the victimizer can be (somewhat) restored, though in different ways, to their pre-crime status as equals. My hope is that by the end of this essay, the reader shall see, contrary to the assumptions of much of the philosophical tradition, that justice and revenge are not antithetical.

## I. FAIRNESS, DEBTS, AND EQUILIBRIUM

Claims about fairness, debt payments, and the restoration of a pre-crime equilibrium are not new to the discussion of punishment. Retributivists often speak of the criminal "owing a debt" or "having to pay for his crime." But these metaphors belong to the vocabulary of restitution and do not cohere very well with the language of retribution. Inevitably, a tension will result because retributivism is concerned with primarily giving the wicked what they deserve, punishing them just because they intentionally did something wrong. Given this emphasis, it is no surprise that retributivists who speak of "debts to be paid" will not do justice to such restitutionist concepts. When they mention debts in their writings, they often leave it unclear who the creditors are or the nature of the currency in which they are to be paid.

We shall see shortly that Herbert Morris[2] is one well-known retributivist who uses the language of debt but fails to locate the main creditors or the proper currency for the payment. F. H. Bradley,[3] on the other hand, gives lip service to the language of restitution, but then completely ignores it. Despite his carelessness, since he is the "purer" retributivist, I shall turn first to his ideas in order to highlight a major oversight and shortcoming of retribution. Then I shall explore Morris's work and point out some holes in his quasi-restitutionist account. In filling these gaps, I shall present an alternative, more robust account of how punishment can be considered a debt.

Like most pure retributivists, Bradley pays too much attention to the evil of the criminal's thoughts, intentions, and will, and ignores the evil that is the harmed condition of the criminal's victim. For Bradley, it is not the evil consequences of the criminal act that must be offset by punishment but the evil embodied in the criminal himself. The emphasis just on the evil residing in the criminal, which is somehow "annulled" by the punishment, makes it seem as if Bradley does not believe, or rather overlooks, the real reason that the criminal's evil thoughts, intention, and will are so abhorrent: the actions resulting from these mental events bring considerable harm. Punishment should not lose sight of this harm. That is why it should include the paying of a debt that offsets the crime's evil consequences. If the harm wrought by the crime is kept in mind, the insistence on just removing the embodiment of evil within the criminal will appear quite unsatisfactory. That Bradley is guilty of such a narrow obsession is evident in his frequently quoted panegyric to retribution reprinted below. There he offers a truncated rationale for punishment in which the victim of the crime goes completely unmentioned.

---

[2] "Persons and Punishment," *The Monist*, v, 24 (1968): 477-501.

[3] *Ethical Studies* (New York: Oxford, 1935, 2d ed.), p. 26: "We *pay* the penalty because we *owe* it, and for no other reason" (italics added).

> The wrong cries out for obliteration...if he can remove it it is his business to do so.... The destruction of the wrong, whatever be the consequences, and even if there are no consequences at all, is still a good in itself...punishment is the denial of wrong by the assertion of right, and the wrong exists in the self, or will, of the criminal, his self is a wrongful self, and is realized in his person and possessions; he has asserted in them his wrongful will, the incarnate denial of right; and in denying that assertion, and annihilating, whether wholly or partially, that incarnation by fine, or imprisonment, or even by death, we annihilate the wrong and manifest the right; and this as we saw, was an end in itself, so punishment is also an end in itself (*ibid.*, pp. 27-29).

While it is not obvious exactly how punishment "annihilates the wrong and manifests the right," it is rather clear that Bradley's approach offers little benefit to the victim of the crime. Morris, on the other hand, does not focus only on the criminal's wickedness. He discusses the losses a crime imposes on others and what sort of debts must be paid for the restoration of equality. But, as we shall see, he does not do justice to our intuitions about victims, fairness, debts, and equilibrium.

Morris believes that members of society accept the restraints of the criminal laws so they can obtain the benefits of others' restraint. Individuals do not violate the property and person of others in order to have the same protection reciprocated. Morris insists that, if a man does not restrain himself from wrongdoing while others do, he obtains an unfair advantage. His illicit gains must be offset by punishment. The burden imposed by punishment replaces the burden he wrongfully avoided when he disregarded the law. Morris says the criminal owes a debt to those who did hold themselves in check (*op. cit.*, p. 478).

The reader might wonder why law-abiding citizens should be compensated for restraining themselves? They each have a duty not to harm innocents. If someone else does harm an innocent, why are they (those who restrained themselves) owed anything? Furthermore, even if we assume that it is unfair to them that the criminal did not restrain himself, how does his suffering compensate them for what they have lost out on? They are not receiving any tangible benefits to offset their extra burdens. The answer must be that Morris just believes that since the law abiding have such a strong interest in the fair distribution of burdens, they will just feel better, and thus it could be said that they have been compensated, when someone who dodged his fair share of burdens eventually receives it.

There are a number of other problems with Morris's account. First of all, he seems to overlook the real victim of unfairness—and that is the person directly harmed by the criminal—the person robbed or raped or the like. It is she who has lost the most due to the criminal transgression. Her right to control her own person or property has been violated. Furthermore, the punishment of her victimizer does not restore an equilibrium of benefits and burdens between herself and him. While the criminal receives through his punishment a burden that he unfairly avoided by not restraining himself, and thus comes to have nominally the same burdens as nearly all of the law-abiding citizens, the actual victim of his assault or deceit is still left far worse off than he. She has not received any tangible compensatory payment for the burdens his wrongdoing unfairly placed on her.

Furthermore, Morris seems actually to overlook the real harms criminals bring to the rest of the law-abiding society while wrongly describing as harms certain features of a number of their actions. The latter mistake is due to the fact that there are countless crimes that most people do not want to commit. Few people have a desire to commit treason, beat old ladies senseless, or molest infants. It is not difficult for the majority of people to restrain themselves, for they are not at all tempted by the above and thus they do not suffer an unfair burden when the rare individual, that is, the criminal, exempts himself from the restraints that Morris wrongly seems to believe weighs so heavily on law-abiding citizens.

The real unfairness to the law-abiding citizens is not that they restrained themselves and the criminal did not, but that they must pay a cost for their safety. They pay high taxes so the police can capture wrongdoers. Knowledge of crimes done to others even leads people to buy security systems and insurance for their own homes and property and to restrict the time and manner that they can go certain places. This same knowledge leads them to live with increased anxiety for the safety of themselves and loved ones. And if the law abiding are especially sensitive, it can also be claimed that another harm is done to them when they suffer vicariously with strangers who are crime victims. Given the above-mentioned costs that the law abiding must bear, it would seem that merely bestowing upon any criminal the burdens he avoided provides inadequate compensation.

## II DEBTS FROM THE DESTITUTE

So a debt theory requires not only a greater but perhaps a more tangible payment than that provided for by Morris's account, and, most importantly, one that goes from the criminal to his direct victim, that is, the person he defrauded, attacked, and so on. Given this requirement, the problem for many restitution-minded retributivists, or their "pure" restitutionist cousins, is to explain how punishment can be called a debt payment in those cases where the individual victim or injured public is not receiving any money, goods, or services from the criminal. Many crimes are neither thefts of wealth nor involve a guilty party that is able to compensate his victims economically for the nonfinancial pain he caused them.

It may be argued that the impoverished victim should still have his illicit gains offset by a jail term. Such punishment is often explained not by the standard retributivist maxim 'that the wicked deserve punishment in proportion to their evil', but that 'no one should benefit from committing a crime'. Although this cannot be easily thought of as a debt payment, perhaps one could understand the punishment as a payment of sorts to others—most notably, the direct victim(s)—who have an interest in the reestablishment of fairness. What they are bothered by is not the Morris-like unfairness in which the criminal avoided burdens that they bore; rather, they just find it unfair that a criminal should keep the rewards of his misdeeds, for he does not deserve such pleasures. The direct victim is especially disturbed by someone's illicit gains that came at his expense. In the present discussion of the criminal destitute at the time of his capture, the felon's illicit benefits are only psychological. Any sadistic or otherwise ignoble satisfactions experienced during a crime or later relived in memory, as well as subsequent pleasures that are made possible by the transgression (for example, the joys of drugs, parties, vacations, and so on brought with the now squandered stolen money), could be offset by the psychological discomforts of a harsh confinement. When the pleasures in which the criminal had no right to partake are "subtracted" by the disutility of the punishment, he will be returned to, approximately, the level of utility he was at prior to his crime. The obvious result of this is that the criminal does not benefit from his crime.

But the principle that a lawbreaker be denied any advantages from a crime would not, by itself, always guarantee a prison term. Where the criminal did something that the law abiding had no desire to do, and if he were apprehended in the act or if the rewards of the crime were taken away before he could enjoy them, then since he did not benefit from his crime, materially or psychologically, a jail sentence will not serve to offset any gains. Any principle that leaves us without grounds for incarceration is obviously unsatisfactory.

Even if the criminal did not receive any unfair advantages from his crime, there is no denying, despite the absence of any tangible restitution, that most directly victimized people, and the public at large, care about a criminal's comfort level. Their concern is in seeing it reduced by imprisonment. Moreover, it appears safe to assume that this is not merely because of either the superior deterrent effects that punishment brings or that incarceration is instrumental to the reform of the prisoner. If the motivation were just deterrence and/or reform, then, if for some reason this could be achieved without offsetting the psychological benefits the criminal received from his conduct, the interest in erasing such wrongful gains, or in their absence, just imposing burdens on the criminal, should vanish. But such interests are unlikely to disappear.

So an answer still has to be given regarding why it is that most of us want to do something about the indigent's criminal's level of psychological well-being when he did not benefit materially nor psychologically from or during his crime, and only avoided lawful restrictions that we, the law abiding, did not feel to be at all burdensome. Because of such cases, it so far looks as if a debt theorist does not have any grounds for demanding imprisonment of the destitute. But a debt theorist might protest that the indigent convict's punishment can consist of his being compelled to take part in some kind of public works that benefit all of society, including the direct victim. This will not do any good, however, if we assume the inmate is too frail to labor. And even if the convict could work, in most cases, his lack of skills would not offset the costs of his confinement.[4] More importantly, even if the criminal's labor could pay for his room and board, the victimized are not likely, in most cases, to consider the standard jobs done by prison work crews to be sufficient compensation for their suffering.

---

[4] The problem of the inability of the unskilled to provide financial compensation plagues Randy Barnett's economic libertarian approach to restitution—"Restitution: A New Paradigm of Criminal Justice," reprinted in Gorr and Harwood, pp. 392-401.

Some readers might think that a nondebt, nonreform, nondeterrence-based justification for punishing those who are indigent can be found in Joel Feinberg's[5] account of expressivism. My contention, however, is that the expression of condemnation can *accompany* punishment that is *primarily* the paying of a debt, but it is a mistake to see the hardship of punishment serving primarily as a "conventional device for the expression of attitudes of resentment and indignation, and of judgements of disapproval and reprobation" (*ibid.*, p. 98). I find what Feinberg calls the "symbolic significance" of punishment to be insufficient, for it is unable by itself to justify the length and severity of sentences that we believe proper. Feinberg claims, however, that:

> given our conventions, of course, condemnation is expressed by hard treatment, and the degree of harshness of the latter expressed the degree of reprobation of the former; still this should not blind us to the fact that it is social disapproval and its appropriate expression that should fit the crime and not hard treatment (pain) as such. Pain should match guilt only as its infliction is the symbolic vehicle of public condemnation (*ibid.*, p. 118).

Despite what Feinberg says, the guiding principle of his theory seems to render the hardships of punishment gratuitous. He himself admits the hypothetical possibility of a society with an "elaborate public ritual, exploiting the trustiest devices of religion and mystery, music and drama, to express in the most solemn way the community's condemnation...rendering superfluous any further hard physical treatment" (*ibid.*, p. 116). Even though our society lacks such a public ritual, most criminals are quite aware that their fellow citizens resent them and disapprove of their conduct. Even those rare individuals who are unaware of this are unlikely to need to serve twenty years behind bars with the four walls of their prison cells serving as blackboards on which society writes its lessons on condemnation.

Instead of emphasizing the severity of the punishment as an expression of society's condemnation of the criminal, Feinberg can still insist that, if the criminal were not punished, it would express a message that the victim and his rights are not important. Feinberg indeed endorses something like this in his discussions under the headings of "authoritative disavowal" and "the vindication of the law." He writes: "What more dramatic way of vindicating (a man's) violated right can be imagined than to have a court forcibly condemn its violation through the symbolic machinery of punishment" (*ibid.*, p. 104)? But this still will not justify maintaining sentences that are extremely harsh. Society can just as well express its concern for the victim and his value through an outpouring of sympathy, regret, and various types of compensation and assistance. And it should not be thought that the reason we do not pursue this expressivist path is just that it would bankrupt the state. Even if the funds were available, there would still exist the need to punish those who victimized us or others.

---

[5] "The Expressive Function of Punishment," in his *Doing and Deserving* (Princeton: University Press, 1970), pp. 95-118.

So it does not appear that expressivism can justify our desire to impose severe sentences upon the worse criminals. We saw that the retributivist defends harsh sentences by insisting that it is just good that the evil are punished even if the victims do not benefit psychologically or materially from the punishment of the criminals who victimized them. Others, most notably J. F. Stephens,[6] doubt that the interest in retribution can be separated from the desire for revenge. He famously wrote: "The criminal law stands to the passion of revenge in much the same relation as marriage to the sexual appetite" (*ibid.*, p. 99). My own view is *not* that retribution is really revenge, though often disguised.[7] Rather, I object to retribution because it provides an insufficient rationale for punishment.

## III RESTITUTION AND REVENGE

My hope is that we can come to see the punishment of both those who gained from their crimes and those who did not as in some way equivalent to the repaying of a debt that makes the injured members of the public *wealthier* in some sense.[8] But again, what benefit could the harmed party receive when those to be punished are already impoverished? It will sound very strange, even outrageous, but I want now to develop the idea that there is a place for revenge in a restitutionist framework for punishment. This is at odds with a long tradition of moral and legal philosophy which has always started with the virtually axiomatic assumption that justice and revenge are diametrically opposed. Nevertheless, I find it reasonable to raise the question whether a punishment that provides little or no payment of any kind to the victim will be sufficient for most people.[9] There was the normal state of affairs, it was disturbed by a crime, then the criminal's illicit gains, if any, were withdrawn or offset. But this does not restore the status quo, for the deep wound and surrounding scar tissue, along with the pique, petulance, and resentment felt by the victimized, remain untreated.

---

[6] *The General View of the Criminal Law of England* (London: Macmillan, 1863).

[7] A plausible example of nonvindictive retribution is when an adult reluctantly informs the police of the crimes and whereabouts of a beloved family member. He is clearly not getting any pleasure bringing his dear relative to justice.

[8] It should not be thought that a wrongdoer cannot be in debt to a person unless he received some sort of benefit from harming him. The model of a loan and its repayment is here misleading; for a person can destroy the property or peace of mind of another and thus have to make restitution, even though he never obtained any pleasure through the illicit action.

Even if the crime were unsuccessful and the victim experienced little loss other than anxiety (and perhaps the money he would then spend on security measures), the victim would still be left greatly bothered by the arrogance and contempt the criminal showed toward him, for it was these very states of mind that initially cleared the way for the villain's transgression. If the criminal had more respect for his target, he would not have been able to violate the latter's wishes through his attempt. So, even if the attempt failed, it still implied the existence of certain attitudes that would insult and enrage the intended victim.

Of course, it is not just the disrespect the criminal feels for the victim that motivates the latter to retaliate. There is in many cases the actual physical, psychological, or financial suffering that the lawbreaker caused. There exists a primitive need to *get back at* those who hurt us in these ways. Acting on this urge is often followed by a cry of "Now we are even!" My thesis is that the vindictive satisfaction that comes from "getting even" should be considered the debt payment that the victim receives from the criminal.

But this equality does not come about *merely* by removing the gains of the villain as was the case in the earlier discussion of the 'no one should benefit from their crime' maxim or the unfair advantages principle that Morris defends. The vindictive person actually feels much better when he exacts his revenge. Revenge (re)creates an equality between the original victim and original victimizer not just by lowering the well-being of the latter but by raising that of the former back to or near the status he enjoyed *prior* to being victimized. This is the most important equilibrium we seek, even if the satisfaction obtained through the harshest revenge still leaves the original victim somewhat beneath his pre-crime status.[10] When the criminal is lowered to the state the victim was at before the latter's compensation, an equality is restored between the criminal and the victim. Both have suffered equivalent harms. This equilibrium is very important in their relationship, more so than offsetting the illicit gains that merely restore the criminal to his pre-crime status. This equality (or near equality) of harms is necessary for the two parties ever to resume a relationship as fellow citizens, each having a right to a certain minimum of formal respect, privileges as well as duties. If the criminal has not been sufficiently harmed, those he directly (and indirectly) victimized, will not ever be able to deal with his return to civil society. When a debt has been fairly determined and properly paid, that is, the vengeance has been exacted but not in a way that is excessive, both the victim and the victimizer can be (somewhat) restored, though in different ways, to their pre-crime status as equals. And it is this restoration of equality, with all its psychological benefits, that justifies the debt theory. Thus it would be a mistake to understand either the criminal's suffering, the vindictive pleasure the victim obtains from it, or any other debt payment in and of itself as a justificatory primitive, that is, the ultimate principle at the end of a chain of justifications.

---

[9] The victim of even a very wealthy criminal is likely to desire not only a great deal of financial compensation but also that his tormentor experience the discomforts of prison life. The existence of such a desire, assuming it also is defensible, is an additional difficulty for Barnett's scheme.

So the victimized individual and/or society is unlikely to forgive and then accept the return of the criminal as a fellow citizen if it is not psychologically compensated to some extent, emotionally replenished after what it has lost in pleasure, security, wealth, liberty, and respect. Even if the victim is exceptionally forgiving and seeks to drop the charge without a debt payment, the criminal cannot go free. First, this much too quick offering of forgiveness is really such a corrupt form of forgiveness that the debt theory offered here does not need to be modified to account for it. Perhaps a person who forgives too quickly has such low self-esteem he does not believe it was that bad for him to be treated as he was. Or the corrupt motivation for prematurely granting forgiveness is the victim really wants to *forget* rather than *forgive*. The mental discomfort that the crime brought—and perhaps the continued state of anxiety resulting from an ongoing conflict—is too hard to bear, so the victim "forgives" his foe in order to have peace of mind. Second, the criminal cannot financially compensate all of society who must pay (in some sense) for the loss of their "peace of mind" as well as in the conventional sense for their own personal security measures and the criminal and legal systems that pursued the wrongdoer. They thus have the right to relish his justified suffering even if his direct victim immediately "forgives" him.

---

[10] But what if the victim is murdered? How can punishment benefit the dead when they obviously cannot experience anything? Will not the only vindictive benefit be that of his surviving family? Perhaps not. Consider the possibility that a person could have his interests frustrated without consciously experiencing their frustration and thus his level of well-being lowered without his knowing it. Examples of such nonexperiential harms are when one never discovers infidelities of a spouse or that physically distant family members one believed to be thriving really were not. If the reader acknowledges that, if he were in these situations, his life would not actually be going as well as he thought it was, then he believes in nonexperiential harms. If there are nonexperiential harms, then there can be nonexperiential benefits. And if a living person can be harmed or benefitted without knowing it, then there is not any reason why a person cannot be posthumously harmed or benefitted. Thus the murder victim can have his interest in revenge and justice satisfied or denied by what happens after his death. For a more detailed account of nonexperiential and posthumous harm, see Feinberg's *Harm to Others* (New York: Oxford, 1984), pp. 79-89.

Therefore, unlike a debt theory that only emphasizes the direct victim, my theory does not need to bring in the extraneous principle of deterrence to justify punishment where the immediate victim has prematurely "forgiven" the criminal. And since the debt theory does not rely upon deterrence as a justification, people are not used merely as a means for society's benefit. Moreover, this instrumentalist justification for punishment, which many find distasteful, can be avoided without a loss in deterrence. The debt theory can indirectly deliver about the same amount of crime prevention as we get if we understand the present system to be a prevention-based one of using men as examples to deter others.

So allowing the satisfaction of vindictive feelings serves as a deterrent and enables the victim and society to accept the criminal's return once they have received their "just fill." Now this "just fill" should not be thought of something separate from punishment, which is how we currently understand compensatory damages through which society benefits the victim. Unlike most restitutionists, I do not distinguish punishment from restitution. I do not believe, as Randy Barnett[11] advocates, that restitution should replace the system of punishment nor that restitution should be added as something over and above punishment. In my theory, punishment and restitution are not separable like a jail sentence and a fine but are one and the same. The satisfaction that the victim derives from the incarceration of his victimizer can be the full payment of the debt—or if accompanied by financial damages, it can still be considered a form of compensation in and of itself.

---

[11] In Gorr and Harwood, p. 393.

## IV. THE PSYCHODYNAMICS OF REVENGE AND RESTITUTION

Perhaps in the best of all worlds, people would not have vindictive feelings. If our natures were different, so would be our system of punishment. But since people do have vindictive feelings, it is prudent to channel them in a productive, civilized way. There already seems to be within men and women a desire for such *judicial vindictiveness* to restore the moral symmetry that existed between individuals prior to a divisive injury. The egalitarianism of this passion is evident in the claim often accompanying revenge: "Now we are even!" This judicial vindictiveness can be distinguished from a type of unjust vindictiveness in which a sadistic delight is experienced when a man is punished more severely than he should be and/or his deprivation is not accompanied by a recognition of his right to someday be restored to civil society as an equal.

If judicial vindictiveness did not exist, a public infected by the other type of vindictiveness would prefer the criminal to find his every day in jail as unbearable as possible. But if the wrongdoer acknowledges his behavior to indeed be misconduct, sees his punishment as a necessary debt, a token offering to his victims that is also essential to his own expiation, then his stay behind bars becomes more tolerable than it would be otherwise. In my estimation, most aggrieved parties would prefer to see such remorse in the felon. They wish him severely disadvantaged but, nonetheless, desire him to agree to his confinement. Were it *only* pain that they sought him to undergo, there would be no interest in his serving an appropriate sentence or recognition of his right to be eventually restored to society.

Some readers might have experienced a desire parallel to what I am calling judicial vindictiveness in their interpersonal exchanges outside of the legal system. If injured in some way, they did not permit the restoration of relations until their victimizers had "paid" some type of nonstandard debt in the form of misery, guilt pangs, or even by being the recipient of a melodramatic slap to the face. Only then arose the feeling that the wrongdoer had suffered enough and could be forgiven and restored, or at least brought closer to his previous status.

On the other side of the equation, the psychodynamics of contrition may be such that the wrongdoer will feel relieved when his victims or their grieving families are satisfied by harsh justice or even revenge being served. Many readers have probably felt at some time much better when after a transgression they suffered some form of sanction, perhaps even bodily pain when they were youngsters, which enabled their spirits to improve as a parallel improvement occurred in the mental states of those who demanded the punishment. In old movies, this often takes the form of one macho but contrite fellow allowing the party he earlier wronged to punch him squarely in the nose or jaw to even the score. The guilty often feel less so when their own suffering satisfies those whom they had earlier victimized. They atone for the crime by accepting such hardships.[12] The debt theory can be considered a form of atonement because (1) the payment provides the criminal an opportunity to make amends for his misconduct; (2) his willingness to make good on his debt supplies evidence that he feels remorse; and (3) the reduction of what he owes his victim can also reduce the pangs of his conscience.

We can even speak sensibly of the criminal having a right to be punished in a culture where punishment is seen as a guilt-removing debt—though in most cases, of course, the criminal would waive this right. The right to be punished is a rather paradoxical way of talking initiated by G. W. F. Hegel though hinted at by Immanuel Kant. They both felt that not punishing the criminal would be to overlook his status as a free and rational citizen whose will is embodied in the law. He was at liberty not to commit the crime, so to withhold punishment is to ignore his free agency and to show him disrespect.[13]

This still is not the full story, however, behind why a man might want to be punished. In the atonement/debt version, what is stressed is the criminal's right to pay a debt so he can be forgiven and restored to the community as an equal. His willingness to accept the punishment reveals his genuine regret, and this makes him more likely to be accepted when he is eventually released. If a guilt-stricken criminal turns himself in to the authorities and demands punishment, it is more likely that he has in mind the atonement form of a right to punishment than the demand that his freedom and rationality be respected, as advocated by Kant and Hegel.[14] He intends through the paying of a debt to lessen his own guilt and the hostility that the society at large feels toward him. His wish is to belong once again to a community of equals, which he realizes cannot be fulfilled unless he is punished.

---

[12] This acceptance of punishment reveals an advantage the atonement/debt theory has over approaches that stress the reform of the criminal. If reform were the raison d'être of punishment, then, at least in principle, some men who commit heinous crimes could avoid jail or be released after only a brief imprisonment, if shortly after their crime they undergo a dramatic change in moral character. This problem of early release plagues Jean Hampton's "The Moral Education Theory of Punishment," in Gorr and Harwood, pp. 356-75. But in the atonement/debt theory, the remorseful prisoner's acceptance of the length of the standard sentence as a debt repayment is *proof* that he is reformed. He atones by willingly serving his full sentence. If he seeks to serve less than the appropriate time, he is attempting to avoid paying his debt in its entirety, which is something a genuinely reformed and remorseful person would not do. So there is neither an epistemological problem of deciding who has been morally transformed and thus should be promptly freed, nor any question of unfairness due to different sentences for the same crime.

[13] Morris argues that we can see that the criminal has a right to be punished on retributivist grounds when we compare this to alternative forms of punishment such as a therapeutic model or a deterrence-based system which could even allow framing innocents. Morris assumes retribution gives a person greater control over his life since he is not harmed except in response to his freely made decisions. But this is questionable. Consider the possibility that framing or treating people lowers the crime rate so much that the probability that any type of harm will befall a man in this hypothetical nonretributivist system is less than it would be in a retributivist society. Thus a man would have more control over his life in a society with such institutions. So Morris's attempt to base the right to punishment on the greater autonomy is potentially self-defeating.

A debt theory that allows the satisfaction of vindictive feelings to assuage the victim's sense of having been wronged even sheds some light on the much discussed problem of moral luck.[15] I have in mind here not the bad moral luck of having a certain disposition or being in certain circumstances others avoided.[16] Rather, I am concerned with the luck that determines whether one is punished harshly for a successful crime or more leniently for a merely attempted but unsuccessful crime.

It strikes many philosophical observers as odd, for example, that the man charged with attempted bodily harm is treated more leniently than the man arrested for successfully committing the same crime, though both villains harbored the same evil intentions. How can the retributivist account for punishing attempts less severely than successful crimes?[17] In both cases, the person is just as evil. Furthermore, if we adopt this retributivist rationale purged of any vindictive payback for actual harm, then, in principle, if we could read minds in a high-tech possible future, we might be obligated to punish those who would like to commit crimes but do not ever begin to set such events in motion for they fear punishment or cannot see how to overcome other solely practical obstacles to the success of the desired crime. I say this because I do not see how most retributivists can draw a line distinguishing those with such evil thoughts, but who are inactive, from those who act on the same type of thoughts.[18] The former are just as evil as those who actually break the law thinking they can avoid capture.[19] It is just nonmoral, contingent factors that keep such people from attempting crimes. And, likewise, it is just nonmoral, contingent factors that keep attempted crimes from being successful.

---

[14] But it should not be thought that the two particular forms of the right to be punished are incompatible.

[15] See Bernard Williams, "Moral Luck," in his *Moral Luck: Philosophical Papers, 1973-1980* (New York: Cambridge, 1981), pp. 20-29; and Thomas Nagel, "Moral Luck," in his *Mortal Questions* (New York: Cambridge, 1979), pp 24-39.

[16] For a list of different ways in which "the natural objects of moral assessment are disturbingly subjected to luck," see Nagel, pp. 28-29

[17] The advocate of deterrence cannot argue that the distinction between successful and unsuccessful attempts came about or is justified by the belief that different sentences will motivate a criminal to stop before his attempt is concluded or will be an incentive not to shoot again after first missing the intended victim. In the latter case, if the stiffer penalty for success did not deter in the first place, it is unlikely to prevent a second attempt where the criminal has even *more* incentive to try again. The additional motivation is that he is more likely to be apprehended and convicted since his unsuccessful attempt leaves alive witnesses to his misconduct. Given the criminal's even greater incentive to make a second attempt than a first, the "logic of deterrence" suggests that the penalty for the second attempt—even if unsuccessful—should be harsher than the penalty for a *successful* first attempt. But we surely are not going to be comfortable with a system that allowed success on the first try to be punished less severely than two-time failures. For instance, the families of murder victims will probably be repulsed by the idea that the killers of their loved ones should be punished less severely than other criminals who have done less harm because they have failed in their attempts to kill someone.

I do not believe, however, that most people care enough about the intrinsic qualities of others' mental states to punish those with evil thoughts just because their thoughts are evil. Or, at least, they do not care to punish them as severely as those who are no more evil but do more harm.[20] So it is doubtful that a retributivist principle like that of Bradley's, which calls for the "annihilation of the evil" embodied within a corrupt man, can provide a good account of not only our actual practices, such as punishing failed attempts less severely, but of our attitudes to counterfactual scenarios.

The leeway given to vindictive feelings provides a way, however, to explain the asymmetrical punishment. The man who commits a violent crime is punished more severely than someone who tried but failed to commit the same crime, since the greater harm the former is responsible for evokes a greater need in the victim, her friends, relatives, and the society at large to retaliate in order to even the score. That is, since they were hurt more, they need more compensation.[21]

---

[18] Only Morris-style retributivists can justify punishing those who are evil *and* act on their evil, while refraining from punishing those who are equally evil but do not act. This is because the latter do not forsake the burdens of restraint and obtain unfair advantages with their criminal attempts.

[19] Moreover, why not entrap the wicked on retributivist grounds? This question was raised—perhaps only in partial seriousness—by Frances Myrna Kamm in conversation.

[20] If some people would insist upon punishing the inactive possessors of these distasteful mental states, it may be because many of them want to retaliate for the animosity, indifference, or some other form of disrespect they directed toward them. This is a form of revenge and not retribution.

[21] It is worth noting that, traditionally, the area where moral luck has bothered us the least has been in matters of compensating victims. For instance, if two *equally* mean-spirited vandals break into different galleries and each destroys an artwork whose value they were unaware of, it does not strike us as unfair that the one who unwittingly destroyed a more valuable painting must make a greater debt payment.

## BOOK REVIEWS

*Autonomous Agents: From Self-control to Autonomy.* ALFRED R. MELE. New York: Oxford University Press, 1995. viii + 271 p. Cloth: $55.00.

The perennial nature of debates about human action, akrasia, and free will are such that one suspects either that everthing of value has been said on the issue or that we should question the parameters and background presuppositions within which they are conducted. Alfred Mele admirably surpasses the former limitation while making suggestive remarks which, if elaborated, might lead to the latter. His goals in this book are straightforward: to give an account of the ideally self-controlled agent, an account that accepts akrasia as a genuine possibility; to argue that additional conditions must be added to those of an ideally self-controlled agent for her to be autonomous; to lay out sufficient conditions for autonomy; and to argue that, with slight variations, such an account can be consistently adopted by both compatibilists and libertarians. His arguments are clear and his strategy perspicuous (even if his prose becomes a bit tortured in the service of precision and detail). I am not sure he is successful at each of these endeavors, but his efforts surely provide a worthy addition to debates that have been thought by some to have atrophied long ago.

Mele's first enterprise is to give an account of intentional action that allows "decisive best judgment" to figure prominently. The problem he addresses concerns how explanations of intentional deliberative action can retain a conceptual or causal tie between judgment and intentional action but allow for cases where we act against our decisive better judgments. Mele's contribution to this puzzle is to argue that the phenomenon of *self-control* figures systematically in intentional action, and akrasia occurs when intervening acts of self-control fail to occur. Hence, even if judgment is understood in the motivational (rather than purely intellectual) sense, akratic action is possible because there are times when efforts at self-control fail to intervene between the judgment of what's best and the competing (possibly stronger) desire to do otherwise (chapters two-four). Mele also applies this model of self-control as an intervening factor in intentional action to cases of belief formation, where agents must overcome a variety of common cognitive biases, as well as processes of emotional control (though Mele by no means identifies "best judgment" with "reason") (chapters five-six). With these arguments in hand, Mele constructs a picture of the ideally self-controlled person.

0022-362X/99/9602/95–99

Perfect self-control is global and reaches overt actions, mental actions, beliefs, emotions, and the like; it is always exercised in the service of decisive, nonakratically held, better judgments; it is exercised whenever appropriate; and it is always effective.

The question, however, is whether such an ideally self-controlled person is autonomous. Mele's answer is that she is not (necessarily). Despite being perfectly able to appraise critically and to act upon her desires and values, the perfectly self-controlled person may nevertheless be acting on desires and values that have been implanted into her by artificial means, ones whose development has bypassed normal reflection and awareness. In those cases, the person is self-controlled but not autonomous (121-26).

Some might argue that autonomy is achieved as long as an agent is fully competent in turning her practical judgments into intentional action (though a fairly robust conception of "competence" might be developed on that score). Mele calls this an *internalist* conception. He argues, however, that there are external conditions required for autonomy. He motivates this by imagining an otherwise fully competent deliberator who is able to act rationally upon values and desires she has critically reflected upon, but whose values have been inculcated in her by a process of brainwashing undertaken against her will. If such a person is less than fully autonomous (and Mele claims that she obviously is), it must be in virtue of some factor external to the inner workings of her decision making and reflection, in particular, relative to the etiology of the values upon which she reflects and acts. Moreover, Mele argues that heteronomy of this sort does not require that the processes of compelled manipulation be intentionally imposed on the agent (163-69).

Mele goes on to argue that what must be true of the history of value formation (or any other characteristic relevant to autonomy) is that the person was not *compelled* to adopt the value or pro-attitude (at least not compelled in a way that she herself arranged). To be compelled is to be unable to shed the attitude in question and to have been conditioned to adopt it in ways that bypass the usual kinds of mental control normal agents enjoy—basic critical reflective capacities as well as skills in self-control.

Many will not be persuaded by Mele's historical argument, perhaps for reasons similar to ones he puts forward himself in criticizing another historical theory of autonomy: aspects of the manner in which one developed one's values are one thing, but one's occurrent ability to act competently on them is another, and only the latter truly matters in the kinds of reactive attitudes, attributions of responsibility, and the like that are the marks of autonomy. But, moreover, it is still unclear that the autonomy-defeating conditions Mele sets out—being compelled in a way that bypasses normal mental control to assess desires in light of reasons—do not obtain so generally that it makes autonomy in this sense an impossibility (especially if they need not be intentionally put in place for autonomy to be disrupted). Mele discusses the education of children, but argues that only in a narrow range of cases are the child's capacities to consider choices in light of reasons *entirely* bypassed, and in those cases we can say the child is robbed of an autonomy-producing upbringing. But we are all raised in relatively limited surroundings, in that we were not, could not have been, given access to information about radically *alternative* ways of life. Mele argues, in fact, that we can lack autonomy when information that would crucially effect our decisions is intentionally withheld (180). So it is unclear how any of us is autonomous, given that, in many ways, everyone's development takes place without full information about what it would be like to live in radically alternative environments (information or experiences which, if pursued, may have changed our values at the core). Consider reports by people who had the chance to live in another culture as a youngster and report how their values were radically changed as a result. What might establish the authenticity of our (necessarily) limited developmental experiences is that we do not retrospectively disapprove of them. But that is equivalent to a condition of historical autonomy that Mele, for other reasons, rejects (138-39).

Mele's initial formulation of autonomy is clearly a compatibilist one, in that all of the processes enumerated in the model can be causally determined, as long as they do not amount to compulsion (not arranged by the agent). But if an incompatibilist libertarian were to insist that there *must* be some indeterministic element in the etiology of intentional acts for them to count as free (autonomous), then Mele has something to offer her as well (chapter twelve). To establish this indeterminacy, Mele focuses on that crucial juncture in decision making where certain beliefs "come to mind" during the process of deliberation. This is the indeterministic link that can occur amidst an otherwise deterministic process of deliberation and choice (captured by the compatibilist conditions of autonomy already described). On this view, a person has proximal control over how she deliberates given the beliefs she has in mind, and more than one physically possible future is open to her during deliberation due to this doxastic indeterminacy. So Mele presents alternative conceptions of autonomy—one compatibilist and one libertarian—and this allows him to maintain an agnosticism about the final truth of compatibilism (251-54).

In the end, however, I am not convinced that the picture of indeterminacy Mele offers will do much to help the libertarian cause, and indeed it may show the crucial fissure in that position. His conception of "control" required by libertarian accounts of freedom (211) amounts to the requirement of *some* indeterminacy regarding action, establishing the central incompatibilist requirement that for freedom to obtain, more than one physically possible future must be possible for a person at the time of deliberation. But indeterminacy does not add up to control. Without a theory of agent causation (which Mele rightly thinks is implausible), it is utterly mysterious how an indeterminate process—whether in the formation of beliefs or desires or whatever—can amount to the kind of self-control required for freedom. If the explanation for my action contains an element of randomness concerning belief formation, for example, then it does not matter, from my perspective, what the source of that belief is; it occurs to me "out of nowhere," as it were. But if that is so, it would not matter if that belief were placed in my deliberative path by a clever manipulator (and if one insists on an element of indeterminacy, just imagine that the manipulator merely arranges a high probability that I shall come upon the surreptitious belief). Since I certainly lack the relevant control in the latter case, why do I retain it in the case of Mele's doxastic indeterminacy?

Mele ends his book with a response to those incompatibilists who, unlike their libertarian cousins, deny the very possibility of autonomy (chapter thirteen). In response to an argument of this sort by Richard Double, Mele argues that merely because there is interminable debate about the nature of a concept or the details of its conditions, as there is with concepts like "freedom," that does not mean that such phenomena do not exist: consider the interminable debates about the nature of intentional action—we certainly do not conclude in that case that there is no such thing as intentional action (239-43).

These observations begin to reveal the limitations of such detailed theorizing about the conceptual structure of complex ideas as a method of solving long-standing metaphysical conundrums. Mele masterfully winds his way through the complex pathways of this conceptual modeling, but in the end it may well be necessary to see concepts like "autonomy" and "freedom" as fully embedded in historically grounded social practices rather than free-floating elements subject to skillful dissection. Indeed, as Mele points out in his discussion of arguments for the impossibility of autonomy, it is often unclear *which* conception (of "autonomy," "responsibility," or the like) a theorist should choose in showing the metaphysical incompatibility of that concept with other (equally malleable) concepts. One may just have to admit that the key notions in these debates may well be too blurry at the edges, and too variable across the different institutional and social settings in which they are used, for any final determination concerning their compatibility to be made apart from analyses of those settings themselves.

JOHN CHRISTMAN

Pennsylvania State University

*A Philosophy of Mass Art.* NOËL CARROLL. New York: Oxford University Press, 1998. xii, 425 p. Cloth $77.00, paper $18.95.

With his new book, Noël Carroll intends to baptize mass art—art produced and distributed for consumption at multiple and discrete sites of reception—as a respectable subject for systematic aesthetics. Toward this end, he exhaustively surveys the entire range of topics related to mass art which might be of philosophical interest. After opening chapters on twentieth-century antagonists and friends of mass art, Carroll advances his own definition of the concept, to which I shall turn below. Through a dialectical engagement with the competing theories advanced by David Novitz and John Fiske, Carroll then spins out the ontological and epistemological implications of his position. Finally, his definitional work completed, Carroll adds three chapters applying recent research on the relation between art and, in order, the emotions, morality, and politics, to mass art.

As this summary indicates, *A Philosophy of Mass Art* is a large-scale undertaking, ambitious both in providing an overview of the state of philosophical work on its subject and in pointing out future directions for inquiry. At the heart of Carroll's project, however, there is a puzzling piece of advice to his fellow aestheticians: we *ought* to take this program seriously because we *ought* to pay more attention to mass art. This is puzzling because Carroll also claims, quite reasonably, that these days all people are aswim in a sea of mass art and so cannot help experiencing it. If we add the likely premise that aestheticians are people, too, this second claim entails that we are al-

0022-362X/99/9602/99–105

ready paying attention to mass art all the time—and if that is true, then Carroll's advice seems otiose.

Of course, Carroll intends the claim that mass art is the dominant art of our time to be a justification for the claim that aestheticians ought to be interested in it: we ought to pay more attention to it because it is the aesthetic air we breathe. This, however, sounds like the stalker's claim to his victim that she ought to be interested in him because he phones so often. This sort of justificatory argument cannot be right, because at least part of what it means to say that mass art (like the stalker) is everywhere is that we, even we philosophers, already respond to it on its own terms just by opening our eyes or ears. Carroll's 'ought', which proposes that we assume *also* a theoretical interest, has a hard time gaining a purchase just because of the experience over which it seeks to exercise command.

This problem can be brought out vividly by considering Carroll's definition of mass art:

> X is a mass artwork if and only if 1. x is a multiple instance or type artwork, 2. produced and distributed by mass technology, 3. which artwork is intentionally designed to gravitate in its structural choices (for example, its narrative forms, symbolism, intended affect, and even its content) toward those choices that promise accessibility with minimum effort, virtually on first.contact, for the largest number of untutored (or relatively untutored) audiences (196).

The pivotal, controversial claim here is that mass art offers its meanings without requiring much in the way of education and thus is, tendentially, art for everyone. Indeed, Carroll even argues that mass art gravitates toward pictorial media because our responses to pictures, unlike to our native languages, are biologically hardwired, requiring no education, strictly speaking, at all (192). But if the defining intention of mass artworks is to render reflection on their reception unnecessary, does it not follow that they seek to render philosophy unnecessary? Even more strongly: if mass art fulfills its "promise" in virtue of rendering idle the kinds of critical and philosophical reflection demanded by complex artworks, would having a philosophy of them not get in their way?

As Carroll well knows, getting in the way of mass art has been the typical philosophical response to it. His opening chapter discusses the views of Dwight MacDonald, Clement Greenberg, R. G. Collingwood, and T. W. Adorno, only to find them all wanting in consistency, cogency, or both. For Carroll, what joins these figures together so that they may be dispatched as a group is that, despite the widely various sources of their supposed failings, they all offer "philosophical resistances to mass art." Carroll's grouping thus conflates two separate points, however, for it suggests that MacDonald et alia fail both because their individual arguments are bad and because they try to resist mass art's promise of easy accessibility. That it is the second of these failures that constitutes Carroll's significant charge against this group is signaled by his identifying a collective "inability to come to terms with mass art" over and above the individually inadequate arguments. Resisting mass art, though, if it is a second "failure" in addition to the mounting of bad arguments against it, is an affective failure which might be better named "refusing to swallow." Unless Carroll can show that gagging is an *improper* response, however, his advice to take mass art seriously amounts to the paradoxical encouragement to swallow what is, by his own definition, designed to be easy to swallow.

In order not to presume that the gag reflex is a philosophical error, Carroll argues that resistance to mass art is a consequence of a deep and widely shared philosophical mistake. Specifically, he avers that a misplaced Kantianism is common to all the resisters and that this theoretical idol makes them unable to come to terms with mass art. According to Carroll, Collingwood et alia use "fragments of the Kantian theory of free beauty to presuppose a theory of art proper" relative to which mass art is judged bad art or not art at all (105); this, he continues, is a two-fold mistake because (a) "Kant had no theory of art at all" (89) and (b) Kantian theories of art are "alien or obsolete frameworks" which are therefore "utterly insensitive to the fact of mass art" (108). Even leaving aside that these two reasons prima facie contradict one another (which may be only a rhetorical slip on Carroll's part) and that poor Collingwood gets peculiarly squeezed into generic Kantianism, this argument still has little to recommend it as a genealogy of the resistance to mass art. What, for instance, is one to make of the claim that Kant had no theory of art? On the presumption that Carroll has not forgotten sections forty-three through fifty-three of the *Critique of Judgment* in which Kant develops a rich theory of art, perhaps he ought to be interpreted as saying overemphatically that Kant had a *bad* theory of art. And perhaps he did, but Carroll's claim that he had none at all gets in the way of his saying why. Instead, we get the observation, delivered in the peremptory voice of the guardian of the dustbin of history, that Kantianism is obsolete and alien. This, however, reduces the argument to circularity: the judgment that the theory is obsolete is supported by the judgment that holders of it cannot come to terms with mass art, while the judgment that what the holders are doing is not coming to terms is supported by the judgment that the theory is obsolete. Still, Carroll uses this circularity as his basis for concluding that the resisters of mass art are insensitive to it. Whether Greenberg, Adorno, or Collingwood were insensitive is an opinion best evaluated by individual readers, but that each dogmatically set himself against accepting art designed to be accessible on the basis of a misplaced, even while nonexistent, Kantian theory of art is an argument that rests on premises and arguments that are unsound.

Even if Carroll could demonstrate that the taste of the resisters was informed by wrong philosophical commitments, however, his claim that resistant taste amounts to insensitivity to mass art begs the question whether mass art really is easy to swallow or, put in Carroll's terms, whether what is designed to be accessible is therefore really accessible. If it is, then resistance to it would be a species of nonresponsiveness, of dogmatic taste, regardless of whether such resistance is motivated by bad philosophy. There is at least one good reason, however, to doubt that mass art is simply accessible, which has as a consequence that there is at least one good reason to think resistance to it is the proper response. We can begin to bring this reason to light once we notice that Carroll's claim that resistance is a perverse response presumes that an artwork (or any artifact) designed to be accessible without much or any specialized background knowledge and thus designed to be used without reference to such knowledge will, if its design is well executed, be familiar and easy to operate. This, though, is a nonsequitur. After I learned to program my video cassette recorder, along came VCR+ with its promise of easy operation; nonetheless, I was befuddled by it because its easy access features were novel relative to my already developed videotaping skills. The new technology was inaccessible to me just because it was designed to be accessible. Similarly, a friend of mine who rarely reads airport novels found *The Bridges of Madison County* impenetrable. In the absence of characters with psychological depth or a theme worthy of literary exploration, she complained, there was no way into the book as an instance of any literary genre with which she was familiar. Granted that the absence of these features was designed for easy access to the story (which, of course, she understood just fine), what was written was still, relative to a specific set of reading skills, inaccessible and unfamiliar.

The reason Carroll does not notice this problem is that he infers from the fact that certain skills of, say, reading, such as following a plot, are widely or even universally distributed, that a novel designed to draw on those skills alone will be accessible (think of this as his Esperanto problem). But this inference skips over the fact that in the absence of an industry intending to cash in on the naked universality of universally distributed skills, mass artworks would be utterly unfamiliar and so hard even to recognize. What we swallow easily is not what is designed to be accessible, but the familiar, and these two concepts do not mean the same thing.

It is a second puzzle that *A Philosophy of Mass Art* elides this distinction since Carroll has equipped himself with all the tools needed to make it. One of the first points he makes in defining mass art is to urge us not to confuse it with popular art; indeed, he plants his flag deeply on this matter by writing that "perhaps the most misleading way to label the phenomenon is to call it *popular* art" (185). If by popular art we mean the "art of the common people," then, he argues, popular art has always existed, whereas mass art, on the other hand, is specific to "modern industrial mass society." I agree that this is an important distinction. If we recall, however, that by Carroll's own definition mass art is also an art of the common people in the sense that it draws on universally distributed capacities, then we ought to infer that mass art is the popular art of modern industrial mass society. This conclusion, though, contradicts the very distinction Carroll insists on. He tries to resolve this problem by specifying the audience for mass art as a mass audience that crosses "national, religious, political, ethnic, racial, and gender boundaries." This, however, just is the distinction between the familiar and the designed-to-be-accessible, since an art that is consumable by its intended audiences only by means of the disavowal of all the vehicles of human identity would be, almost by definition, an art unfamiliar to most humans. Now, if Carroll imported this distinction between the familiar and the designed-to-be-universally-accessible back into the fundamental distinction between popular and mass, then he could thematize mass art as a kind of production and distribution which aims to establish historically novel alternatives both to popular art and to the common people; this, in turn, would allow him to block the problematic conclusion that mass art is our popular art. Oddly, though, Carroll refuses this saving strategy by arguing instead that the difference between the audiences for the two is merely numerical: popular art is liked by "lots" of people, mass art by "lots and lots" of people (187). This is clearly a distinction without a difference which effectively erases the distinction between mass art and popular art. Instead, Carroll could and should have crafted a real distinction by considering how the mass audience intended by mass art is an audience that is conceptually and practically different *in kind* from the culturally differentiated audiences presupposed by popular—familiar—arts.

Of course, it may turn out (or already have turned out) that this distinction between mass art and popular art gets erased as a matter of historical fact, that what is designed to be accessible will become the most familiar. For that to happen, though, two other historical changes in the audience for art will have to happen also. First, adult human beings will have to undergo a process of infantilization whereby they develop a taste for what is designed to be readily understood. I call this infantilization because it returns us to our early aesthetic pleasures, such as sing-songy poems and stories of mighty heroes; through those simple pleasures we developed skills that would be used later in meeting more challenging experiences, but as mass art becomes familiar art we are uncannily confronted by our childish pleasures, recalling us to our more primitive skills, in the guise of adult entertainment. Second, human beings will have to develop a distaste for what will have been familiar art previously. When partisans of local cultures complain about the Americanization of global culture, they are complaining about mass art (the production of which is overwhelmingly American) driving out the familiar arts of diverse locales. Looked at from the nonmassified, that is, local and locale-specifying, context, that mass art is designed to be accessible means that it is designed to make otherwise familiar and popular forms inaccessible. Carroll rightly claims that mass art and avant-garde art are the alternative forms of artistic modernity, but this leads him to think that therefore avant-garde art is the privileged antagonist of mass art. In fact, though, while from its own perspective the avant-garde may be locked in a contest with mass art, mass art is locked in a contest with the popular arts which it must render unfamiliar if it is to turn its own design imperative of accessibility into familiarity. Mass art, in short, is not accessible in virtue of being designed to be; rather, it becomes accessible by driving out the ability to respond to art that demands reflection. It is this historical dynamic to which the philosophical resistances to mass art respond in their setting philosophy hopefully against mass art (and if aesthetics is not up to the job, so much the worse for it).

If I am right here, then my earlier reference to the stalker was more than an analogy. Mass art manifests the same logic as the stalker: they are both designers of the accessible which insist that they are thereby our familiars. But just as it would be an addled piece of advice to a stalker's victim that she should "come to terms" with the stalker, so, too, it is strange in the case of mass art. In the first case, we would advise the victim to figure out effective ways to keep the stalker away—to keep him unfamiliar despite his design imperatives—for the sake of a life lived better in virtue of not being absorbed, theoretically or otherwise, in false and damaging offers of intimacy. So, too, for systematic aesthetics, the resistance to mass art, not despite its design imperatives but because of them, is the proper response in the sense that it is the proper way not to "come to terms" with it. Its insinuating familiarity, calculated, as Carroll argues, to render reflection unfamiliar, earns it our odium.

GREGG M. HOROWITZ

Vanderbilt University

---

The Department of Philosophy
University of California, Riverside
and the
University of California
Humanities Research Center
present

UCR's 9th Annual Philosophy Conference
**Debating Essentialism**
March 26 and 27, 1999

Speakers: John Dupré, Eli Hirsch, Elisabeth Lloyd, Ed Stein, Charlotte Witt, and Naomi Zack

For registration and conference information, please contact Gmarti@citrus.ucr.edu, or sschauer@pop.ucr.edu. Visit our website at http://www.ucr.edu/philosophy/conference99.html.

## NEW BOOKS: ANTHOLOGIES

ALANEN, LILLI, SARA HEINÄMAA, and THOMAS WALLGREN, eds.: *Commonality and Particularity in Ethics.* New York: St. Martin's, 1997. x, 493 p. Cloth $55.00.

ALCOFF, LINDA MARTÍN, ed.: *Epistemology: The Big Questions.* Malden, MA: Blackwell, 1998. xiii, 445 p. Cloth $62.95, paper $32.95.

ALEXANDER, LARRY, ed.: *Constitutionalism: Philosophical Foundations.* New York: Cambridge, 1998. ix, 319 p. Cloth $54.95.

ALLEN, RICHARD, and MURRAY SMITH, eds.: *Film Theory and Philosophy.* New York: Oxford, 1997. xi, 474 p. Cloth $75.00.

ALPERSON, PHILIP, ed.: *Musical Worlds: New Directions in the Philosophy of Music.* University Park: Penn State UP, 1998. vi, 188 p. Paper $14.95.

ARIEW, ROGER, JOHN COTTINGHAM, and TOM SORELL, eds.: *Descartes'* Meditations: *Background Source Materials.* New York: Cambridge, 1998. xviii, 270 p. Cloth $54.95, paper $18.95.

BACON, FRANCIS: *The History of the Reign of King Henry VII and Selected Works.* BRIAN VICKERS, ed. New York: Cambridge, 1998. xlv, 284 p. Cloth $59.95, paper $19.95.

BARNES, ANNETTE, ed.: *Seeing through Self-Deception.* New York: Cambridge, 1997. x, 182 p. Cloth $54.95.

BAR ON, BAT-AMI, and ANN FERGUSON, eds.: *Daring to Be Good: Essays in Feminist Ethico-Politics.* New York: Routledge, 1998. xix, 252 p. Cloth $75.00, paper $21.99.

BATTIN, MARGARET P., ROSAMOND RHODES, and ANITA SILVERS, eds.: *Physician Assisted Suicide: Expanding the Debate.* New York: Routledge, 1998. 463 p. Paper $21.99.

BECHTEL, WILLIAM, and GEORGE GRAHAM, eds.: *A Companion to Cognitive Science.* Malden, MA: Blackwell, 1998. xvii, 791 p. Cloth $99.95.

BENDER, THOMAS, and CARL E. SCHORSKE, eds.: *American Academic Culture in Transformation: Fifty Years, Four Disciplines.* Princeton: University Press, 1998. xii, 371 p. Paper $16.95.

BERNARD-DONALS, MICHAEL, and RICHARD R. GLEJZER, eds.: *Rhetoric in an Antifoundational World.* New Haven: Yale, 1998. 468 p. Cloth $35.00.

BIGONGIARI, DINO, ed.: *The Political Ideas of St. Thomas Aquinas.* New York: Free Press, 1997. xxxvii, 215 p. Paper $15.95.

BILETZKI, ANAT, and ANAT MATAR, eds.: *The Story of Analytic Philosophy: Plot and Heroes.* New York: Routledge, 1998. xiv, 276 p. Cloth $70.00.

BLOCK, NED, OWEN FLANAGAN, GÜVEN GÜZELDERE, eds.: *The Nature of Consciousness.* Cambridge: MIT, 1997. xxix, 843 p. Paper $29.95.

BOHMAN, JAMES, and MATTHIAS LUTZ-BACHMANN, eds.: *Perpetual Peace: Essays on Kant's Cosmopolitan Ideal.* Cambridge: MIT, 1997. 260 p. Paper $17.50.

BOHMAN, JAMES, and WILLIAM REHG, eds.: *Deliberative Democracy: Essays on Reason and Politics.* Cambridge: MIT, 1997. xxx, 447 p. Cloth $45.00, paper $25.00.

BOLINGBROKE, HENRY ST JOHN: *Political Writings.* DAVID ARMITAGE, ed. New York: Cambridge, 1997. xliv, 305 p. Cloth $64.95, paper $24.95.

BONTEKOE, RON, and MARIETTA STEPANIANTS, eds.: *Justice and Democracy: Cross-Cultural Perspectives.* New York: Cambridge, 1997. 477 p. Cloth $44.95, paper $12.95.

BOUCHER, DAVID, ed.: *The British Idealists.* New York: Cambridge, 1997. xlvi, 304 p. Cloth $64.95, paper $24.95.

BROCK, GILLIAN, ed.: *Necessary Goods: Our Responsibilities to Meet Others' Needs.* Lanham, MD: Rowman and Littlefield, 1998. ix, 239 p. Cloth $63.00, paper $23.95.

BROWNING, DOUGLAS, and WILLIAM T. MYERS, eds.: *Philosophers of Process.* New York: Fordham, 1998. xiv, 449 p. Cloth $35.00, paper $19.95.

BURNYEAT, MYLES, and MICHAEL FREDE, eds.: *The Original Sceptics: A Controversy.* Indianapolis: Hackett, 1997. xiii, 155 p. Cloth $34.95, paper $14.95.

BYRNE, ALEX, and DAVID R. HILBERT, eds.: *Readings on Color, Volume I: The Philosophy of Color.* Cambridge: MIT, 1997. xxviii, 317 p. Cloth $30.00; *Readings on Color, Volume II: The Science of Color.* Cambridge: MIT, 1997. xxxii, 465 p. Cloth $30.00.

CALHOUN, CRAIG, ed.: *Hannah Arendt and the Meaning of Politics.* Minneapolis: Minnesota UP, 1997. viii, 362 p. Cloth $54.95, paper $21.95.

CARRUTHERS, PETER, and JILL BOUCHER, eds.: *Language and Thought: Interdisciplinary Themes.* New York: Cambridge, 1998. xiv, 324 p. Cloth $59.95, paper $19.95.

CASATI, ROBERTO, and CHRISTINE TAPPOLET, eds.: *European Review of Philosophy, Volume 3: Response-Dependence.* Stanford: CSLI, 1998. vi, 277 p. Cloth $59.95, paper $22.95.

CHANG, RUTH, ed.: *Incommensurability, Incomparability, and Practical Reason.* Cambridge: Harvard, 1998. xii, 303 p. Cloth $57.50, paper $24.95.

CHAPPELL, VERE, ed.: *Descartes's Meditations: Critical Essays.* Lanham, MD: Rowman and Littlefield, 1997. xvi, 272 p. Cloth $34.00, paper $12.95.

CLARK, ANDY, and PETER MILLICAN, eds.: *Connectionism, Concepts, and Folk Psychology: The Legacy of Alan Turing, Volume II.* New York: Oxford, 1996. ix, 281 p. Cloth $55.00.

CLARK, KELLY JAMES, ed.: *Philosophers Who Believe: The Spiritual Journeys of Eleven Leading Thinkers.* Downers Grove, IL: Intervarsity, 1997. 284 p. Cloth $55.00.

CLARKE, SAMUEL: *A Demonstration of the Being and Attributes of God and Other Writings.* EZIO VAILATI, ed. New York: Cambridge, 1998. xxxvii, 168 p. Cloth $59.95, paper $18.95.

COLAPIETRO, VINCENT M., ed.: *Reason, Experience, and God: John E. Smith in Dialogue.* New York: Fordham, 1997. 158 p. Cloth $28.00, paper $18.00.

COLEMAN, JULES L., and CHRISTOPHER W. MORRIS, eds.: *Rational Commitment and Social Justice: Essays for Gregory Kavka.* New York: Cambridge, 1998. viii, 253 p. Cloth $54.95.

CRIGGER, BETTE-JEAN, ed.: *Cases in Bioethics: Selections from the Hastings Center Report.* New York: St. Martin's, 1998. xxiv, 295 p. Paper $24.72.

CRITCHLEY, SIMON, and WILLIAM R. SCHROEDER, eds.: *A Companion to Continental Philosophy.* Malden, MA: Blackwell, 1998. xv, 680 p. Cloth $84.95.

CULLITY, GARRETT, and BERYS GAUT, eds.: *Ethics and Practical Reason.* New York: Oxford, 1997. viii, 421 p. Cloth $77.00.

D'AGOSTINO, FRED, and GERALD F. GAUS, eds.: *Public Reason.* Brookfield, VT: Ashgate, 1998. xxiii, 470 p. Cloth $153.95.

DAVIDSON, ARNOLD I., ed.: *Foucault and His Interlocutors.* Chicago: University Press, 1997. 261 p. Cloth $30.00.

DAVIS, STEVEN, ed.: *Philosophy and Theological Discourse.* New York: St. Martin's, 1997. xiv, 232 p. Cloth $59.95.

DE BARY, WM. THEODORE, and TU WEIMING, eds.: *Confucianism and Human Rights.* New York: Columbia, 1998. xxiii, 327 p. Cloth $40.00.

DEUTSCH, ELIOT, ed.: *A Companion to World Philosophies.* Malden, MA: Blackwell, 1997. xviii, 587 p. Cloth $89.95.

DEWEY, JOHN: *The Essential Dewey, Volume I: Pragmatism, Education, Democracy.* LARRY A. HICKMAN and THOMAS M. ALEXANDER, eds. Bloomington: Indiana UP, 1998. xiii, 417 p. Cloth $45.00, paper $27.95; *The Essential Dewey, Volume II: Ethics, Logic, Psychology.* LARRY A. HICKMAN and THOMAS M. ALEXANDER, eds. xii, 425 p. Cloth $45.00, paper $27.95; *Principles of Instrumental Logic: John Dewey's Lectures in Ethics and Political Ethics, 1895-1896.* DONALD F. KOCH, ed. Carbondale: Southern Illinois UP, 1998. xii, 192 p. Cloth $39.95.

DIQUINZIO, PATRICE, and IRIS MARION YOUNG, eds.: *Feminist Ethics and Social Policy.* Bloomington: Indiana UP, 1997. xv, 302 p. Cloth $39.95, paper $17.95.

DOMAŃSKA, EWA, ed.: *Encounters: Philosophy of History after Postmodernism.* Charlottesville: Virginia UP, 1998. xii, 293 p. Cloth $55.00, paper $19.50

DUFF, ANTHONY, ed.: *Philosophy and the Criminal Law.* New York: Cambridge, 1998. vii, 261 p. Cloth $54.95.

ELGIN, CATHERINE Z., ed.: *The Philosophy of Nelson Goodman, Volume I: Nominalism, Constructivism, and Relativism in the Work of Nelson Goodman.* New York: Garland, 1997. xvii, 278 p. Cloth $77.00; *The Philosophy of Nelson Goodman, Volume II: Nelson Goodman's New Riddle of Induction.* New York: Garland, 1997. xvii, 290 p. Cloth $77.00; *The Philosophy of Nelson Goodman, Volume III: Nelson Goodman's Philosophy of Art.* New York: Garland, 1997. xvi, 364 p. Cloth $88.00; *The Philosophy of Nelson Goodman, Volume IV: Nelson Goodman's Theory of Symbols and Its Applications.* New York: Garland, 1997. xviii, 324 p. Cloth $88.00.

THE JOURNAL OF PHILOSOPHY

VOLUME XCVI, NO. 3, MARCH 1999

UNIVERSITY LIBRARY
PUR
CAMPUS

# OF CONSPIRACY THEORIES*

"The only thought which philosophy brings with it, in regard to history, is the simple thought of Reason—the thought that Reason rules the world, and that world history has therefore been rational in its course."
—G.W.F. Hegel, *The Philosophy of History*

"Shit happens." —Popular contemporary bumper-sticker slogan

The millennium is nigh, and with each passing year, the American consciousness is increasingly in the grip of conspiratorial thinking.[1] Some conspiracy theories are the stuff of legend. Every year best-selling books are published, block-buster movies produced, and high-rated television and radio programs aired which seek to convince us that Lee Harvey Oswald did not act alone in the assassination of John F. Kennedy; that, in 1947, an alien spacecraft crashed near Roswell, New Mexico, and the United States govern-

---

* I wish to thank my philosophical colleagues at Washington University/St. Louis, where this paper was originally presented. Special thanks to David Hilditch, Pim Haselager, Pete Mandik, Jim Moore, and Chase Wrenn for insightful comments and discussion.

[1] Conspiracy theory has not been given much attention by philosophers. In fact, I am aware of only a handful of discussions: for example, Karl Popper, *The Open Society and Its Enemies*, Volume 2: *The High Tide of Prophecy. Hegel, Marx, and the Aftermath* (London: Routledge, 1966, 5th ed.), pp. 94-99; and Charles Pigden, "Popper Revisited, or What Is Wrong with Conspiracy Theories?" *Philosophy of the Social Sciences*, XXV (1993): 3-34. I believe that the reason for this omission is that most academics simply find the conspiracy theories of popular culture to be silly and without merit. I believe, however, that it is incumbent on philosophers to provide analysis of the errors involved in common delusions, if that is indeed what they are. I offer this paper in the spirit of Philip Kitcher's work on the philosophical difficulties of scientific creationism—*Abusing Science: The Case against Creationism* (Cambridge: MIT, 1982).

0022-362X/99/9603/109–26

ment recovered the craft and its extraterrestrial occupants, then covered up with stories of crashed weather balloons; that the rapid introduction of crack cocaine into America's urban centers during the late 1980s was facilitated by Central Intelligence Agency-backed, Nicaraguan Contra-affiliated, drug smugglers. This plot benefitted both the CIA (by channelling money to the Contras after Colonel Oliver North's Iranian arms deals had been uncovered) and their status quo seeking supporters (by keeping the economic and cultural boot firmly on the neck of Black America).

Here is a test of your conspiracy literacy. How many of the following conspiracy theories have you heard about before? How many are you tempted to believe have merit?

(1) HIV (human immunodeficiency virus), the virus that causes AIDS (acquired immune deficiency syndrome), was the product of American or Soviet biological warfare research before it was released (either intentionally or otherwise) on an unsuspecting world.
(2) Extraterrestrials regularly visit our planet, mutilating cattle and abducting humans (whose memories are then erased). Our government is aware of this situation.
(3) Los Angeles once had an efficient mass transit system based on street-cars, but in the 1930s and 1940s automakers, rubber manufacturers, and oil companies colluded with city officials there to dismantle this system in order that Los Angeles could become a model city of automobile-based transportation.
(4) Trans World Airlines Flight 800 was accidentally downed by a U. S. Navy missile; a fact then covered up by a government fearful of bad press in an era of post-Cold War military downsizing.
(5) All transatlantic communications are monitored and recorded by the U. S. National Security Agency.
(6) Significant aspects of the world economy are under the control of a small group of individuals, be they Freemasons, the Trilateral Commission, or a secret organization of Jewish bankers.

I could go on. The point of this list is to make clear just how pervasive conspiracy theory now is. Prognostication is a dangerous business, but I predict that future cultural critics and sociologists are going to have as much to say about our contemporary obsession with conspiracies as they now have to say concerning the implications of the 1950s American fascination with Unidentified Flying Objects and extraterrestrial invasions.

## I

The present essay is epistemological, however, not sociological. I contend that the study of conspiracy theories can shed light on the nature of theoretical explanation. Conspiracy theories, as a general category, are not necessarily wrong. In fact, as the cases of Watergate and the Iran-Contra affair illustrate, small groups of powerful individuals do occasionally seek to affect the course of history, and with some nontrivial degree of success. Moreover, the available, competing explanations—both official and otherwise—occasionally represent dueling conspiracy theories, as we shall see in the case of the Oklahoma City bombing.

The definition of conspiracy theory poses unexpected difficulties. There seems to exist a strong, common intuition that it is possible to delineate a set of explanations—let us call them *unwarranted conspiracy theories* (UCTs).[2] It is thought that this class of explanation can be distinguished analytically from those theories which deserve our assent. The idea is that we can do with conspiracy theories what David Hume[3] did with miracles: show that there is a class of explanations to which we should not assent, *by definition*. One clear moral of the present essay is that this task is not as simple as we might have heretofore imagined.

Before continuing, I should emphasize that at no point should the reader conclude that I am giving arguments for or against the truth of any given explanation. The issue here is not whether aliens are indeed visiting our planet, or whether Oswald acted alone. Ultimately, in these cases, there is a historical fact of the matter. These facts are not manifest, however, and we must theorize and speculate as to what has happened. The issue here is one of warranted belief. In other words, it may well be correct that "the truth is out there," but given our epistemic situation, we ought not necessarily believe everything which is, in fact, true. In this respect, we are in the same situation as Hume. As Thomas Huxley[4] observed, Hume cannot say that miracles have never happened, only that, even if they have, we have no warrant to believe them. Hume has no way of determining, with certainty, whether Jesus turned stone into bread and fed the multitude. Maybe He did or maybe he did not. Hume *is* in a position, however, to say whether we *ought to believe* this miracle occurred given the evidence at hand (or even given the *possible* evidence at hand).

Understanding *why* we are not warranted in believing certain conspiracy theories can make clearer why we ought to believe the things that we should. I propose to make a study of UCTs, and in determining where they go wrong, attempt to tell a story about explanation properly conducted. In the following section, I shall turn to a recent event that has spawned a significant number of conspiracy theories: the Oklahoma City bombing. This discussion of a real event and the conspiracy theories (some warranted, some not) will help me illustrate the analysis provided in the rest of the paper. In section III, I shall discuss the problem of trying to define UCTs and illustrate the difficulties for finding analytic criteria for distinguishing good from bad conspiracy theories. I turn to the alleged virtues of UCTs in section IV in order to explain both their current popularity as well as the grounds we have for ultimately rejecting them. Such explanations feature significant degrees of explanatory breadth and are not simply unfalsifiable in nature. But continued belief in a UCT requires more and more pervasive skepticism in people and public institutions. This leads me in section V to a discussion of conspiratorial thinking in the context of competing visions of the nature of the world. I suggest that confronting UCTs forces us to choose between the pervasive skepticism entailed by these theories and an absurdist view of the world entailed by their rejection. I conclude by suggesting that it is philosophy's job to show us the way out of this dilemma.

---

[2] Other, less charitable, readers have suggested such labels as 'kooky', 'weirdo', and 'harebrained', among others. While colorful, I do not find such terminology conducive to serious discussion

[3] "Of Miracles" (section X of *Enquiries concerning Human Understanding*, 1748), reprinted in S. Tweyman, ed., *Hume on Miracles* (Bristol: Thoemmes, 1996), pp. 1-20.

[4] "The Order of Nature: Miracles" (chapter VII of Huxley's *Hume*, 1881), reprinted in Tweyman, pp. 161-68.

## II

To give us a concrete example of conspiracy theory in action, it is necessary to deal with one in some detail. The detail is relevant because it is from the details of such events that UCTs take their start. I hope the reader will excuse this short diversion.

At a few minutes past nine on the morning of Wednesday, April 19, 1995, a Ryder rental truck—parked in front of the nine-story Alfred P. Murrah Federal Building in Oklahoma City, Oklahoma—is torn apart by a powerful explosion. Some 168 people, including—most tragically—nineteen children in the building's day-care center, are killed in what has since been designated the single worst act of terrorism ever carried out on American soil.

As emergency personnel move in to rescue survivors from the rubble, a huge federal investigation is launched. Initial suspicion falls upon possible Middle-Eastern terrorists. This turns out to be a dead-end.[5] After tracing the destroyed truck to Elliott's Body Shop in Junction City, Kansas, attention is soon focused on two white males in their twenties. Composite sketches of "John Doe #1" and "John Doe

---

[5] Although not before Oklahoma resident and American citizen, Abrahim Ahmad, is detained in London, strip-searched, and returned to the United States in shackles. He is quickly ruled out as a suspect and released.

#2" are released to the public on the following day. A nationwide manhunt ensues.

Meanwhile, it happens that at 10:20 am on the morning of the bombing, Officer Charles Hanger of the Oklahoma State Troopers spotted a northbound 1977 Mercury Marquis on Interstate 35 approximately sixty miles north of Oklahoma City. The car was travelling at eighty miles per hour and lacked visible license plates. Officer Hanger pulled the car over without incident, but as he approached the lone driver, he noticed a suspicious bulge under the driver's jacket. The driver volunteered that he was indeed armed, at which point the officer drew his own weapon, pointed it at the head of the driver, and disarmed him of a loaded pistol and a sheathed five-inch bladed knife. The driver, twenty-six-year-old Persian Gulf War veteran Timothy McVeigh, was arrested on charges of transporting a loaded weapon, carrying a concealed weapon, and operating a motor vehicle without license plates. McVeigh was still in a cell awaiting arraignment two days later, when he was identified as "John Doe #1" and linked to the Oklahoma City bombing.

The investigation continued and, some two years later, McVeigh, along with Terry Nichols and Michael Fortier, were convicted in connection with the case. McVeigh and Nichols were tried on a charge of "Use of a Weapon of Mass Destruction," and several other counts, including eleven counts of First-degree Murder for the eleven federal agents killed in the blast. (The government contends that the defendants intentionally set out to kill federal agents.) Fortier plead guilty to lesser crimes in relation to the bombing in return for his testimony against McVeigh and Nichols. Nichols was convicted on manslaughter and conspiracy charges, but the jury refused to convict him of the more serious crime of murder. McVeigh was convicted on all counts and sentenced to death.

In its cases against McVeigh and Nichols, the government contended that McVeigh, Nichols, and Fortier planned and carried out the bombing of the Oklahoma City federal building as an act of terrorism against what they saw as a powerful and dangerous governmental entity. (The government claims that McVeigh and his co-conspirators were inspired in part by A. MacDonald's *The Turner Diaries*,[6] a novel which describes a racial revolution in America and which begins with a small group bombing the Federal Bureau of Investigation headquarters in Arlington, Virginia.) McVeigh, so goes the story, was particularly upset by the deadly federal raid of the Branch Davidian compound in Waco, Texas, and sought to avenge the deaths of these victims of federal heavy-handedness. (The Oklahoma City bombing occurred on the second anniversary of the fiery end of the Waco stand-off.) So, McVeigh (trained in explosives by the U. S. Army) and Nichols constructed a powerful ammonium nitrate and fuel oil bomb in the back of a rented truck and detonated it in front of the Oklahoma City federal building.

---

[6] (Washington, D.C.: National Alliance, 1978; New York: Barricade, 1996, 2nd edition)—latter cited here

The bombing provides us with a nice example of the dynamics of conspiracy theory. Within days of the event, questions were raised about the official account of the bombing (itself calling for a conspiracy): Did McVeigh, Nichols, and Fortier act on their own, or were they a small part of a larger team? Were they set up as "patsies" to take the fall for the crime, as Oswald claimed he was before Jack Ruby gunned him down? Was the bombing perhaps carried out either directly or indirectly by elements within the U. S. government who wished to sway public sentiment toward greater law-enforcement powers and against "far right-wing" ideologues who have been increasingly setting themselves at odds with federal authorities? Many such questions have been raised in the years following the bombing.

The grounds for such conspiracy theories are investigated in a book, *OKBomb! Conspiracy and Cover-up*, by Jim Keith (perhaps best known for his 1995 *Black Helicopters over America: Strikeforce for the New World Order*).[7] Keith does not endorse any particular alternative account, but instead raises many issues and questions for the official account. For example, he casts doubt upon the alleged guilt of the government's prime suspect by listing eleven, "striking incongruities in the behavior of Timothy McVeigh prior to and during his arrest," including the following:

> (1) "Of high curiosity is the fact that McVeigh, even though reported to possess fake I.D. under [other names], chose to give his correct name and address to the owner of the Dreamland Motel he was staying at in Junction City [just prior to the bombing]. This is hardly the behavior of a man planning on committing a crime of the magnitude of the Oklahoma City bombing."[8]
>
> (2) We are to believe that McVeigh was fleeing the scene of the crime in a car without license plates: "All the car seemingly lacked was a big sign saying 'Arrest me!'" (*ibid.*, p. 23).

---

[7] Lilburn, GA: IllumiNet, 1996 and 1995, respectively.

[8] *OKBomb*, pp. 28-29.

(3) "If McVeigh had just taken part in the bombing of the Murrah Building [...], is it likely that he, armed with a pistol, would have let a police officer approach his car without resisting him? As a man trained in the use of small arms, McVeigh would not have had much trouble in gunning the officer down from within the protection of the car" (*ibid.*, p. 30).

(4) When McVeigh was arrested, he gave as his address the property of James Nichols, the brother of Terry Nichols, allegedly the place where the bomb-making materials were stored: "Does it make any sense that McVeigh would have pointed the FBI to his alleged partner or partners in crime..." (*ibid.*, p. 31)?

Keith adds to the above observations the following: there were early news reports suggesting that the Bureau of Alcohol, Tobacco, and Firearms (BATF) had received prior warning of the bombing, a claim supported by the fact that no BATF employees were in the building at the time of the blast, but were on the scene within minutes. The BATF vociferously deny that they were forewarned, but no explanation has ever been given of the early reports. And there is the mysterious "John Doe #2"—never caught and left unexplained by the official story that led to the convictions of McVeigh and Nichols.

The answer to these questions and others raised in Keith's book are intended to lead the reader to the conclusion that all is not as we have been led to believe. Probing beneath the surface of the facts of the case as they have been presented in the mainstream media reveals the real possibility that the Oklahoma City bombing may have not been the act of a lone group of politically-motivated criminals, but rather suggests something much bigger and much more sinister. One popular conspiracy theory related to the bombing goes like this: a group of right-wing ideologues (including McVeigh, Nichols, Fortier, and the mysterious John Doe #2) were indeed plotting to blow up a federal building. Their actions were being monitored by the BATF, however. (On some accounts, their actions were being influenced by the BATF. John Doe #2 was actually a BATF informer, or perhaps even a plant. The BATF hoped that swooping in and stinging a group of "dangerous, right-wing terrorists" at the very last moment would do much to erase their public image as an organization of bumbling incompetents resulting from the fiasco in Waco.) In any case, McVeigh and his friends were involved, but only tangentially. McVeigh helps assemble the bomb, but he is unaware of the exact plans for its use, or is actively misled. At the last moment, the BATF screws up, loses contact with the group or are outsmarted by them and the terrorists successfully carry out their act of terror. McVeigh—unaware that the bombing has occurred—is picked up by the police. The BATF realize that they have a public relations nightmare on their hands: they knew about the bombing, but through sheer incompetence and a desire to grandstand, failed to prevent it. When McVeigh is picked up in an unrelated incident, they see their chance to cover up their own involvement in and knowledge of the incident. He is the perfect patsy because he does have some involvement in the incident, but does not know the whole story.

Notice that this account is not completely without plausibility. The BATF *does* have a public image of incompetence and institutional insecurity. On most accounts, they fouled up the original raid on the Branch Davidian compound because they were more interested in national television coverage than in competent law enforcement. It makes sense to suggest that this is an image they are actively seeking to overcome. Also, this would not be the first time that a federal undercover agent has incited not-so-innocent citizens almost to carry out crimes they might not have without such encouragement. Finally, it certainly would not be the first time that a law-enforcement agency has covered up the fact that it has incompetently allowed a crime to be committed.

## III

What is a conspiracy theory? A conspiracy theory is a proposed explanation of some historical event (or events) in terms of the significant causal agency of a relatively small group of persons—the conspirators—acting in secret. Note a few things about this definition. First, a conspiracy theory deserves the appellation "theory," because it proffers an *explanation* of the event in question. It proposes reasons why the event occurred. Second, a conspiracy theory need not propose that the conspirators are all powerful, only that they have played some pivotal role in bringing about the event. They can be seen as merely setting events in motion. Indeed, it is because the conspirators are not omnipotent that they must act in secret, for if they acted in public, others would move to obstruct them. Third, the group of conspirators must be small, although the upper bounds are necessarily vague. Technically speaking, a conspiracy of one is no conspiracy at all, but rather the actions of a lone agent.

This then might be considered a bare-bones definition of conspiracy theory, be it warranted or otherwise. UCTs—at least the ones that concern me in this essay—have a number of additional characteristics:

(1) A UCT is an explanation that runs counter to some received, offi-

cial, or "obvious" account.

Central to any UCT is an official story that the conspiracy theory must undermine and cast doubt upon. Furthermore, the presence of a "cover story" is often seen as the most damning piece of evidence for any given conspiracy; the fact that someone has gone to such pains to create a false explanation points to a conscious effort to deceive.

(2) The true intentions behind the conspiracy are invariably nefarious.

I am aware of no popular conspiracy theory according to which some group of powerful individuals is secretly doing *good* but desperately hopes its schemes will not be revealed.

(3) UCTs typically seek to tie together seemingly unrelated events.

You might think that the Oklahoma City bombing had nothing to do the 1995 Sarin gas attack in the Tokyo subway. What if I told you, however, that the appropriate anniversary was not the federal raid in Waco (which occurred two years earlier to the day) but rather the next day, one month to the day since the gas attack in Tokyo? The connection is that the attack in Tokyo was carried out by the CIA in retaliation for the Japanese bugging of the Clinton White House, all part of an ever-escalating Japanese-American trade war. So the Oklahoma City bombing was a "pay-back" hit carried out by the Japanese....[9] As will be discussed below, this unifying aspect of conspiracy theories is a key feature of their apparent explanatory strength.

(4) As noted, the truths behind events explained by conspiracy theories are typically well-guarded secrets, even if the ultimate perpetrators are sometimes well-known public figures.

What seems to drive many conspiracy theorists is the deeply held belief that, if only the truth can be publicly revealed (in much the way that Bob Woodward, Carl Bernstein, and "Deep Throat" exposed the Watergate break-ins), the conspirators' nefarious plans will be thwarted. Therefore, it is in the interest of conspirators to see to it that the truth is not revealed, or if it is, that it is not widely believed.

(5) The chief tool of the conspiracy theorist is what I shall call ***errant data***.

---

[9] Let me note here that ***I am not making this stuff up.*** Throughout this paper, the examples I offer all come from sources other than myself. For example, see Keith's *OKBomb!*, chapter 21, for more on the "Japanese connection" to the Oklahoma City bombing.

Errant data come in two classes: (a) unaccounted-for data and (b) contradictory data. *Unaccounted-for data* do not contradict the received account, but are data that fall through the net of the received explanation. They are data that go unexplained by the received account. For example, the early reports that the BATF had prior warning of the Oklahoma City bombing and the fact that no BATF employees were in the building at the time of the explosion represent unaccounted-for data with respect to the received account of the bombing. *Contradictory data* are data that, if true, would contradict the received account. McVeigh's manifest idiocy in fleeing the scene of the bombing in a car without license plates is a contradictory datum with respect to the official account of him as conspiratorial ringleader capable of planning and carrying out such a terrorist operation. (The role of errant data will be discussed in more detail below.)

These criteria go some way toward distinguishing UCTs from conspiratorial explanations that are less epistemically problematic. I am not concerned here with conspiracies to throw surprise birthday parties for friends or attempts by parents to deceive young children about the existence of magical beings.[10] Such day-to-day conspiracies typically do not meet all of these additional criteria. Surprise birthday parties are not organized for nefarious purposes, nor are they meant to be kept secret perpetually. These criteria, however, do not distinguish UCTs from all conspiracies we are warranted in believing. Both Watergate and the Iran-Contra Affair meet all of these criteria, yet belief in these conspiracies seem prima facie warranted.[11] There is no criterion or set of criteria that provide a priori grounds for distinguishing warranted conspiracy theories from UCTs. The philosophical difficulties of UCTs require a deeper analysis, to which I now turn.

## IV

Conspiracy theories are attractive, a fact demonstrated by their current popularity. But their alleged virtues are subtly flawed. I shall show how certain elements of conspiracy theory *appear* to be those of legitimate theoretical explanation, but argue that on closer analysis, they are not. Therefore, I am trying to do two things: (1) explain why it is that UCTs are so popular, but (2) explain why we should nonetheless fail to believe in them. They are popular, I suggest, because they exhibit several well-known explanatory virtues. They are nonetheless unwarranted because they exhibit these virtues in ways that undermine the strength of those virtues.

---

[10] I cannot help but wonder whether conspiracy's grip on Western thought can be blamed, at least in part, on the number of conspiracies we experience as children. As we grow older, we discover just how many things adults have been systematically lying to us about: Santa Claus, the Easter Bunny, the Tooth Fairy, where babies come from, and so on

[11] Some might balk at the "nefariousness" criterion. Presumably, North and his colleagues believed they were serving some higher good by their attempt to circumvent Congress's prohibitions on funding Nicaraguan rebels. I do not wish to get into a debate over the semantics of 'nefariousness' in the Machiavellian context of politics. Suffice it to say that however one characterizes what went on the Iran-Contra Affair and Watergate, they were not completely above board. If they were, secrecy would not have been so necessary.

The first and foremost virtue which conspiracy theories exhibit, and which accounts for much of their apparent strength, is the virtue of unified explanation or explanatory reach. According to this virtue, all things being equal, the better theory is the one that provides a unified explanation of more phenomena than competing explanations. Unified explanation is the *sine qua non* of conspiracy theories. Conspiracy theories *always* explain more than competing theories, because by invoking a conspiracy, they can explain *both* the data of the received account *and* the errant data that the received theory fails to explain. So, for example, in our case of the Oklahoma City bombing, conspiracy theories explain the data of the official story. On the account discussed above, the bombing was carried out by terrorists of which the BATF and other agencies were aware, but, due to sheer incompetence, failed to stop. This theory explains why and how the Murrah building came to be bombed. Furthermore, it explains the various errant data: for example, why no BATF personnel were in the building (they were forewarned) and the bizarre behavior of McVeigh (who was innocent of the crime and hence not expecting to be framed by federal agents in search of a scapegoat).

This is the beauty of conspiracy theories. They offer wonderfully unified accounts of all the data at hand, both those the official story explains, plus those niggling, overlooked errant data. As I shall now try to show, however, UCTs obtain unity of explanation at too high a cost.

The role of errant data in UCTs is critical. The typical logic of a UCT goes something like this: begin with errant facts, such as the observation that no BATF employees were injured in the Oklahoma City bombing and the early reports of prior warning. The official story all but ignores this data. What can explain the intransigence of the official story tellers in the face of this and other contravening evidence? Could they be so stupid and blind? Of course not; they must be intentionally ignoring it. The best explanation is some kind of conspiracy, an intentional attempt to hide the truth of the matter from the public eye.

By invoking a conspiracy hypothesis, large amounts of "evidence" are thrown into question. This is one of the most curious features of these theories: to my knowledge, conspiracy theories are the only theories for which evidence *against* them is actually construed as evidence *in favor* of them. The more evidence piled up by the authorities in favor of a given theory, the more the conspiracy theorist points to how badly "They" must want us to believe the official story.

Let me note two things at this point. First, conspiracy theories are not alone in placing great emphasis on errant data. The history of science is replete with examples of theoretical innovation initiated by an investigation into data that did not fit the standard paradigm. It is a good pragmatic heuristic for scientific effort to be expended on chasing after errant data, in the hopes that these loose strings might lead to the unraveling of currently misguided theory. What conspiracy theories get wrong, however, is that the existence of errant data *alone* is not a significant problem with a theory. Given the imperfect nature of our human understanding of the world, we should *expect* that even the best possible theory would not explain *all* the available data. One's theory should not fit *all* the available data, because not all the available data are, in fact, true.[12] Invariably, some of our measurements, some of our interpretations and other theories get something wrong about the nature of the world.

Second, the problematic of conspiracy theories goes beyond simple false data. If the only problem with UCTs was that they place too much emphasis on small sets of data at odds with an official account, then that would not make them a very interesting phenomenon. Conspiracy theories differ from most other theories in one very interesting way, however. Conspiracy theorists would rightly point out that they have one problem with which scientists are not faced. By hypothesis, the conspiracy theorist is struggling to explain phenomena that other, presumably powerful, agents are actively seeking to keep secret. Unlike the case of science, where nature is construed as a passive and uninterested party with respect to human-knowledge gathering activities, the conspiracy theorist is working in a domain where *the investigated actively seeks to hamper the investigation.* Imagine if neutrinos were not simply hard to detect, but actively sought to avoid detection! This is exactly the case with which conspiracy theorists contend we are confronted in the cases they seek to explain. This is why countervailing evidence and lack of evidence can and ought to be construed as supporting their theories.

---

[12] I have heard this sentiment attributed to Francis Crick.

This brings me to the most commonly voiced complaint about UCTs, namely, that they are simply unfalsifiable. The worry is that given a situation where all potentially falsifying evidence can be construed as supporting, or at worst as neutral evidence, then conspiracy theories are by definition unfalsifiable. In favor of conspiracy theorists, it should be noted that this unfalsifiability is not as ad hoc as it might initially seem, due to the active nature of the investigated, just noted. It is not ad hoc to suppose that false and misleading data will be thrown your way when one supposes that there is somebody out there actively throwing that data at you. Just ask Kenneth Starr. As evidenced by any number of twentieth century, U. S. government-sponsored activities (take your pick), we have reason to believe that there exist forces with both motive and capacity to carry out effective disinformation campaigns.

My claim here is that unfalsifiability is only a reasonable criterion in cases where we do not have reason to believe that there are powerful agents seeking to steer our investigation away from the truth of the matter. Falsifiability is a perfectly fine criterion in the case of natural science when the target of investigation is neutral with respect to our queries, but it seems much less appropriate in the case of the phenomena covered by conspiracy theories. Richard Nixon and North actively sought to divert investigations into their respective activities and both could call upon significant resources to maintain their conspiracies. They saw to it that investigators were thwarted in many of their early attempts to uncover what they accurately suspected was occurring. Strictly hewing to the dogma of falsifiability in these cases would have led to a rejection of conspiracy theories at too early a point in the investigations, and may have left the conspiracies undiscovered.

No, the problem with UCTs is not their unfalsifiability, but rather the increasing degree of skepticism required by such theories as positive evidence for the conspiracy fails to obtain. These theories throw into doubt the various institutions that have been set up to generate reliable data and evidence. In doing so, they reveal just how large a role trust—in both institutions and individuals—plays in the justification of our beliefs. The problem is this: most of us—including those of us who are scientists and who work in scientific laboratories full of expensive equipment—have never carried out the experiments or made the empirical observations that support most contemporary scientific theories. Unless we want to conclude that the vast majority of us are not warranted in believing that the platypus is a mammal and that gold is an atomic element, we need some procedure by which the epistemic warrant obtained by those who *do* make the appropriate observations can be transferred to the rest of us. In modern science, this procedure involves the elaborate mechanisms of publication, peer review, professional reputation, university accreditation, and so on. Thus, we are warranted in believing the claims of science because these claims are the result of a social mechanism of warranted belief production.[13]

In the public sphere where conspiracy theories dwell, there are related mechanisms for generating warranted beliefs. There is the free press, made up of reporters, editors, and owners who compete to publish "the scoop" before others do. There are governmental agencies charged with investigating incidents, producing data, and publishing findings. And there are, of course, various "free agents" (including the conspiracy theorists themselves) who are members of the public. Inherent in the claim that alleged evidence against a theory should be construed as evidence for that theory is a pervasive skepticism about our public, fact-gathering institutions and the individuals working in them. Thus, as a conspiracy theory matures, attempt after attempt to falsify a conspiracy theory appears to succeed, and this apparent success must be explained as the nefarious work of the conspirators. As a result of this process, an initial claim that a small group of people is conspiring gives way to claims of larger and larger conspiracies.

In the case of the Oklahoma City bombing, the initial conspiracy only involved the BATF agents and their immediate superiors. The FBI and other federal agencies were brought into the investigation, yet did not report a BATF conspiracy. Therefore, they must have been brought into the conspiracy. Ditto for certain members of the press who must have stumbled across evidence of the conspiracy, but who have not yet disclosed it in the national media. What began as a small conspiracy on the part of a few members of a paramilitary U. S. federal agency invariably swells into a conspiracy of huge proportions, as positive evidence for the alleged conspiracy fails to obtain. And, as more people must be brought into the conspiracy to explain the complicity of more and more public institutions, the less believable the theory should become.

---

[13] For more on the critical role of testimony in the epistemic process, see C. A. J. Coady, *Testimony. A Philosophical Study* (New York: Oxford, 1992). For more on aspects on the social construction of warranted belief, see Helen Longino's *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry* (Princeton: University Press, 1990), and "The Fate of Knowledge in Social Theories of Science," in Frederick F Schmitt, ed., *Socializing Epistemology: The Social Dimensions of Knowledge* (Lanham, MD: Rowan and Littlefield, 1994), pp. 135-57; and Kitcher's "Socializing Knowledge," this JOURNAL, LXXXVIII, 11 (November 1991): 675-76, and "Contrasting Conceptions of Social Epistemology," in Schmitt, pp. 111-34.

It is this pervasive skepticism of people and public institutions entailed by some mature conspiracy theories which ultimately provides us with the grounds with which to identify them as unwarranted. It is not their lack of falsifiability per se, but the increasing amount of skepticism required to maintain faith in a conspiracy theory as time passes and the conspiracy is not uncovered in a convincing fashion. As this skepticism grows to include more and more people and institutions, the less plausible any conspiracy becomes.

Consider another famous UCT, the one claiming that the Holocaust never occurred and is a fabrication of Jews and their sympathizers. Robert Anton Wilson correctly notes that "a conspiracy that can deceive us about 6,000,000 deaths can deceive us about *anything*, and that it takes a great leap of faith for Holocaust Revisionists to believe World War II happened at all, or that Franklin Roosevelt did serve as President from 1933 to 1945, or that Marilyn Monroe was more 'real' than King Kong or Donald Duck."[14] In the process of holding onto a belief in an increasingly massive conspiracy behind more and more public events, we undermine the grounds for believing in anything. At some point, we shall be forced to recognize the unwarranted nature of the conspiracy theory if we are to left with *any* warranted explanations and beliefs at all.

## V

I want to take a step back and get a handle on the broader philosophical issues raised by conspiracy theories and the implication of their current surge in popularity. I contend that conspiracy theories embody a thoroughly outdated world view, a perspective on the meaning of life that was more appropriate in the last century. Recognizing this anachronistic element of conspiratorial thought is useful, however, if it reveals something about the contemporary Zeitgeist. Furthermore, the present popularity of conspiracy theories suggests that we are now in the grip of a conflict between world views.

Conspiracy theorists are, I submit, some of the last believers in an ordered universe. By supposing that current events are under the control of nefarious agents, conspiracy theories entail that such events are *capable of being controlled*. In an earlier time, it would have been natural to believe in an ordered world, in which God and other supernatural agents exercised significant influence and control. With the rise of materialist science and capitalist economies—peaking in the eighteenth and nineteenth centuries—the notion of an ordered universe was still held to, but the role of the supernatural was either greatly diminished (as in Deism) or eliminated (as in Marxism). As G. W. F. Hegel puts it in the passage quoted in the epigraph, "world history has...been rational in its course." Therefore, on this view, there is some hope that humans can understand, predict, and conceivably control the course of human events. This the conspiracy theorists believe, only they further believe that the wrong folks are at the helm.

---

[14] "Beyond True and False: A Sneaky Quiz with a Subversive Commentary," in Ted Schultz, ed., *The Fringes of Reason* (New York: Harmony, 1989), pp. 170-73, here p. 172.

Such beliefs are out of step with what we have generally come to believe in the late twentieth century. The rejection of conspiratorial thinking is not simply based on the belief that conspiracy theories are false as a matter of fact. The source of the problem goes much deeper. The world as we understand it today is made up of an extremely large number of interacting agents, each with its own imperfect view of the world and its own set of goals. Such a system cannot be controlled because there are simply too many agents to be handled by any small controlling group. There are too many independent degrees of freedom. This is true of the economy, of the political electorate, and of the very social, fact-gathering institutions upon which conspiracy theorists cast doubt. Even if the BATF were part of a large conspiracy to cover up their incompetence in the Oklahoma City bombing, it is implausible to believe that not a single member of the BATF stationed in Oklahoma would be moved by guilt, self-interest, or some other motivation to reveal that agency's role in the tragedy, if not to the press, then to a lover or family member. Governmental agencies, even those as regulated and controlled as the military and intelligence agencies, are plagued with leaks and rumors. To propose that an explosive secret could be closeted for any length of time simply reveals a lack of understanding of the nature of modern bureaucracies. Like the world itself, they are made up of too many people with too many different agendas to be easily controlled.

The rejection of the conspiratorial world view, however, is not something about which I am particularly thrilled. If conspiracy theories are genuinely misguided, then I fear we are left with an apparently absurdist image of the world. A lone gunman can change the course of history when the U. S. President just happens to drive past the window of his place of work during the gunman's lunch hour. The conspiratorial world view offers us the comfort of knowing that while tragic events occur, they at least occur for a reason, and that the greater the event, the greater and more significant the reason. Our contemporary world view, which the conspiracy theorist refuses to accept, is one in which nobody—not God, not us, not even *some* of us—is in control. Furthermore, the world (including the people in it) is uncontrollable, irrational, and absurd in a way illustrated by the plays of Eugene Ionesco and Samuel Beckett.

It is not that the rejection of conspiratorial thinking entails that the world is *random*, only that it is without broad meaning and significance. One can reject conspiracies about Kennedy's assassination and still hold that Oswald's behavior was caused and fully deterministic in ways on which science can get a grip. Such a scientific understanding of Oswald's behavior is absurdist, however, in just the sense I intend here. There was no human or rational agency responsible for his behavior on such account, only the blind workings of brain chemicals and childhood emotional traumas. Events like Kennedy's assassination and the Oklahoma City bombing have had an enormous emotional and meaningful impact on our world; indeed, they represent singularities after which American culture has never been the same. Conspiracy theories have as a virtue the attempt to preserve a human meaning—a rational accounting—for these sea changes which allows them to be understood in human terms. If the conspiracy theorists are wrong, then when considered from the human perspective, all we seem to be able to say is that "Shit happens."

Considered in this light, the challenge of conspiracy theory is that it forces us to choose between an almost nihilistic degree of skepticism and absurdism: the conspiracy theorist chooses to embrace the hyperskepticism inherent in supposing dissimulation on a truly massive scale (by distrusting the claims of our institutions) over the absurdism of an irrational and essentially meaningless world. Until a third option is presented—and perhaps this is one of the jobs of philosophy—we should expect UCTs to continue enjoy significant popularity.

## VI

What lessons should we take away from these musings on the dynamics of conspiracy theory? The story I have told here seems to point to three main, related morals.

First, the folly of conspiracy theories highlight the fact that we should not be overly bothered when our theories of social events fail to make sense of all the data. Invariably, not all the data are true. In particular, the irrational and fallible nature of humans should lead us to expect that some of the data generated by us are certainly wrong. Witnesses misremember the past or exhibit unconscious biases. Reporters and government agents will get things wrong in the early moments of a crisis, and will later be loathe to admit those mistakes. For this reason, a theory that has as one of its main features a unified account of all of the data in a variety of seemingly unrelated occurrences should be called into question on those grounds alone. We expect our explanations to be good, but we also expect them to be imperfect.

Second, we should be careful not to over-rationalize the world or the people that live in it. Rejecting conspiratorial thinking entails accepting the meaningless nature of the human world. Just as with the physical world, where hurricanes, tornadoes, and other "acts of God" *just happen*, the same is true of the social world. Some people just do things. They assassinate world leaders, act on poorly thought out ideologies, and leave clues at the scene of the crime. Too strong a belief in the rationality of people in general, or of the world, will lead us to seek purposive explanations where none exists.

The third lesson of conspiracy theories is that we ought to recognize such theories as embodying an almost nihilistic degree of skepticism about the behavior and motivations of other people and the social institutions they constitute. To the extent that a conspiracy theory relies on a global and far-reaching doubt of the motives and good will of others, it is akin to global philosophical skepticism. These extreme skeptical stances should be dealt with in the same way. We should be wary of theoretical accounts that expend more energy undermining the epistemic warrant of competing explanations than on generating new, positive evidence.

So, in the end, what do I think of conspiracy theories? My initial motivation was to present an analysis of conspiracy theories in the spirit of Hume's analysis of miracles. For Hume, miracles are *by definition* explanations that we are never warranted in believing. If my analysis here is correct, however, we cannot say the same thing about conspiracy theories. They are not *by definition* unwarranted. (A good thing given that we want to believe in at least *some* conspiracies—for example, Watergate and Iran-Contra.) Instead, I suggest that there is nothing straightforwardly analytic that allows us to distinguish between good and bad conspiracy theories. We seem to be confronted with a spectrum of cases, ranging from the believable to the highly implausible. The best we can do is track the evaluation of given theories over time and come to some consensus as to when belief in the theory entails more skepticism than we can stomach. Also, I suspect that much of the intuitive "problem" with conspiracy theories is a problem with the *theorists* themselves, and not a feature of the theories they produce. Perhaps the problem is a psychological one of not recognizing when to stop searching for hidden causes. Nonetheless, I suggest that the study of conspiracy theories, even the crazy ones, is useful, if only because it forces us clearly to distinguish between our "good" explanations and their "bad" ones.

BRIAN L. KEELEY

Washington University/St. Louis

# THE EMPEROR'S NEW INTUITIONS

One of the favorite argumentative methods of present-day analytic philosophers is to appeal to intuitions. I argue here that such appeals are usually without any respectable theoretical foundation. Moreover, the best way of justifying the use of intuitions in philosophical argumentation entails radical changes in our ways of thinking about intuitions.

Where does the current popularity of appeals to intuition come from? The timing of the great revival of intuitionist methodology gives us a clue to its causes. Before the early 1960s, you could scarcely find any overt references, let alone appeals, to intuitions in the pages of philosophical journals and books in the analytical tradition. After the mid-1960s, you will find intuitions playing a major role in the philosophical argumentation of virtually every article or book. Why the contrast?

The answer is simple. Intuitions came into fashion in philosophy as a consequence of the popularity of Noam Chomsky's linguistics and its methodology. According to a widespread conception, generative linguists like Chomsky were accounting for competent speakers' intuitions of grammaticality by devising a grammar, that is, a set of generative rules that produces all and only such strings that are intuitively accepted by these speakers. This kind of methodology was made attractive by the tremendous perceived success of Chomsky's theories in the 1960s and 1970s. Not only was transformational grammar the *dernier cri* in linguistics, it was seen as a major revolution in the study of language. What is more, it was taken to provide a methodological paradigm of what can be done in those fields where the subject matter involves the tools of human thought and cognition. The use of intuitions in philosophical argumentation thus originated from philosophers' attempt to get on the bandwagon of transformational grammar.

This bandwagon effect was amplified by the example of some well-known philosophers. An influential case in point was Saul Kripke.[1] As early as 1963-1964, he was led by his "natural intuition" to the view "that the names of ordinary language are rigid designators" (*ibid.*, p. 5). Furthermore, he acknowledges himself that his argumentation for this view rests on intuition: "In these lectures, I will argue, intuitively, that proper names are rigid designators..." (*ibid.*, p.

---

[1] *Naming and Necessity* (Cambridge: Harvard, 1980), p. 5.

0022-362X/99/9603/127–47

49). Kripke also argues as if his main problem were the correct interpretation of these intuitions; for instance, whether they can be explicated so as to pertain to the relative scopes of different logically active notions (*ibid.*, p. 10), or whether Michael Dummett or Kripke himself has the right view on "linguistic intuitions." Quite obviously, Kripke helped to give the intuitionist bandwagon a mighty push. Unfortunately, his reliance on intuitions in defending his idea of rigid reference is apt to give intuitions a bad name. I shall return to his use of intuitions later.

In reality, the linguistic parentage of contemporary philosophers' intuitionist methodology nevertheless constitutes a strong reason to be wary of it. Ironically, the attribution of such a mode of argumentation to Chomsky is at best a half-truth. Speakers' intuitions first entered into his thinking as an antidote to oversimplified views concerning the ways in which a grammar could, in principle, be constructed for a previously unstudied language from a corpus of observed utterances. One of the goals of some of his linguistic predecessors had been what Chomsky called *decision procedure*, that is, a systematic method of choosing among different grammars, each of which purports to account for a given corpus of utterances. He argues against the feasibility of such a method and, among other things, adduces intuitions as alternative grounds of comparisons among different grammars. According to John Lyons:[2]

> In *Syntactic Structures*[3] he [Chomsky] says that the sentences generated by the grammar should be "acceptable to the native speaker," and he considers that it is a point in favor of the kind of grammar he develops that it also accounts for the "intuitions" of native speakers with respect to the way certain sentences are recognized as equivalent or ambiguous (*op. cit.*, pp. 32-33).

Indeed, the first explicit proposal of an intuition-based grammar is not found in Chomsky but in the famous (or notorious) review of *Syntactic Structures* by Robert B. Lees,[4] who elevates one's "feeling for language"—or at least feeling for the syntax of language—to the status of "the empirical basis of grammatical analysis":

> ...as we see from the practice of linguists, though unfortunately not from their own descriptions of linguistic methodology, the criterion actually used in all crucial cases is either the informant's response in carefully designed...elicitation techniques, or else the linguist's own

---

[2] *Noam Chomsky* (New York: Penguin, 1977, revised edition).

[3] Mouton: The Hague, 1957; see pp. 49-50.

[4] *Language*, XXXIII (1957): 375-407

> Sprachgefühl is called upon to provide correct analysis, after which any ad-hoc rule may be devised to designate the results (*ibid.*, p. 309).

Lees spells this out further in a footnote:

> It is precisely this Sprachgefühl, this *intuitive* notion about linguistic structure, which together with the sentences of a language, *forms the empirical basis of grammatical analysis*; and it is precisely the purpose of linguistic science to render explicit and rigorous what ever is vague about these linguistic feelings (*ibid.*, p. 309, fn. 40; emphases added).

Any linguist or any philosopher who is familiar with the subsequent argumentative practice in her field can only have here a feeling of *déjà vu* all over again. What is going on is obviously that Lees was trying to align Chomsky's methodology with that of empirical sciences (such as Lees's own first *métier*, chemistry), with intuitions playing roughly the same role as empirical observations.

Chomsky's willingness to countenance Lees's construal of his methodology amounted to a significant change of mind on his part; for Chomsky's original strategy did not involve appeals to intuition. In *The Logical Structure of Linguistic Theory*,[5] he writes that a linguistic

> ...theory is significant just to the extent that we can demonstrate that the data do have the prescribed form, that is, to the extent that we have unambiguous, cross-culturally valid tests for applying the undefined notions to data. Thus if one of the basic undefined terms of linguistic theory is "intuition," and if we define phonemes in this theory as elements which our intuition perceives in a language, than the notion of phoneme is as clear and precise as is "intuition." ...It should be clear, then, why the linguist interested in constructing a general theory of linguistic structure, in justifying given grammars or (to put the matter in its more usual form) in constructing procedures of analysis should try to avoid such notions as "intuition" (*ibid.*, pp. 86-87).

Admittedly, Chomsky himself later moved closer to an intuitionist methodology. In his later work, according to Lyons, "Chomsky includes intuitions of the speakers of a language as part of the data to be accounted for by the grammar" (*op. cit.*, p. 33). Chomsky also quickly qualified the role of intuitions in the methodology of linguistics by introducing the distinction between *competence* and *performance*. But this distinction is highly questionable and, predictably, it did nothing to cool philosophers' ardor in their appeals to intuitions. For by the token of Chomsky's distinction, to hesitate in one's appeals to one's intuitions would have been ipso facto to doubt one's own linguistic or, more generally, conceptual competence.

---

[5] New York: Plenum, 1975.

What is there to be said about philosophers' toeing the Chomskian party line? The most amazing fact about the current fashion of appealing to intuitions is the same as the proverbial dog's walking on two feet: not that it is done particularly well but that it is done at all. For what is supposed to be the justification of such appeals to intuition? One searches the literature in vain for a serious attempt to provide such a justification.

This blind faith is below the intellectual dignity of philosophers for whom unexamined intuitions should not be worth intuiting. It is also in stark contrast to the procedure of earlier philosophers. Every major philosopher or philosophical tradition whose argumentation relied on intuitions (or their equivalent) provided an explanation why, according to their lights, intuitions could provide the right kind of information or insight.

For Aristotle, for instance, to think of *X* is to realize the form of *X* in one's soul. This instantiation of the form is quite as authentic as an instantiation outside the soul. The instantiated form in the soul has the same powers and the same necessary links to other forms as an instantiation outside the soul. Hence I can find all about this form "intuitively" simply by realizing it in my mind. As one might put it, thought experiments for Aristotle were literally real experiments. Because of this, the immediate awareness of what there is in my soul, called by Aristotle *noûs*, could give me the first premises of a science.

It is hard to think of a better justification for appeals to intuition. Contemporary philosophers do not believe in the reality of forms, however, no matter whether they are supposed to be instantiated in the mind or in the world, nor have any earlier philosophers done so since the days of the medieval nominalists. Our contemporaries do not have the courage of their intuitions and claim infallibility for them, as Aristotle did. Their use of intuitions as prima facie evidence has nothing to do with Aristotle's *noûs*. If they want to find an Aristotelian precedent to their methodology, the obvious candidate is not Aristotle's reliance on *noûs*, but his use of reasonable common opinions, *endoxa*, as starting points of his dialectical arguments. The resemblance between them and twentieth-century philosophers' "intuitions" is uncanny, when it comes to the role they are supposed to play in philosophical argumentation. But I doubt that anyone before Kripke would have called such hunches "intuitions."

But even many postnominalist philosophers had a theory or theories to back up their appeals to intuitions. The best-known form of such theories is the doctrine of innate ideas of the rationalists. An enterprising historian of ideas might try to trace this line of defense of intuitive knowledge to Plato's *anamnesis*. Be that as it may, it was the doctrine of innate ideas that provided the rationale for appeals to intuition (whether under this title or not) by rationalists like René Descartes.

When both Aristotelian forms and innate ideas became discredited, no respectable justification for intuition as a separate source of insights remained. Accordingly, by the eighteenth century the notion of "intuitive knowledge" became watered down and ended up meaning little more than immediate knowledge. When John Locke cites as a paradigmatic example of intuitive truths that one plus two equals three, he is not postulating a separate mental faculty of intuition or even a separate source of knowledge. This minimal sense of "intuition" has played a much more important role in subsequent philosophical terminology than is generally realized. For instance, interpreters like Emmanuel Levinas[6] who see in Husserlian *Anschauung* a special kind of "theoretical act of consciousness that makes objects present to us" might be well advised to consider the ancestry of Edmund Husserl's notion.

Part of this ancestry is Immanuel Kant's theory of how intuitions, even in a minimal sense of immediate representations of particulars, can yield synthetic knowledge a priori. His explanation turns on the idea that in sense perception we impose a certain relational structure on experience, namely, the structure of space and time. Hence we can recover the relations we have ourselves imposed on our experience by reproducing them in imagination. What is remarkable here is that this Kantian justification of the use of intuitions in mathematics does not presuppose any particular mental faculty of intuition nor any sense of "intuition" stronger than Kant's own definitory *unmittelbare Vorstellung von dem Einzelnen*.

Now, what conceivable theoretical rationale do contemporary philosophers' appeals to intuitions have? The embarrassing answer is: none of the above. The vast majority of philosophical writers these days take the name 'intuition' in vain since they do not believe in Platonic anamnesis, Aristotelian forms, Cartesian innate ideas, or Kantian transcendental deductions. Nor can they find aid and comfort in the other, minor source of allegedly intuitionist methodology in twentieth-century philosophy. G. E. Moore's argumentation is often labeled "intuitionistic." If so, the term 'intuition' is used in the eighteenth-century minimal sense in which the phrase 'intuitive knowledge' is merely an euphemism for 'immediate knowledge'. Moore's method is predicated on his belief that in any experience we can distinguish the immediately given object of experience from the experience itself as a subjective happening that is merely part of one's consciousness. His appeals to "intuition" are therefore in reality appeals to what is objectively given to me in immediate experience. Not only is such a notion of "intuitively given" independent of any assumption of a special mental faculty of intuition, it presupposes that there is no such special faculty. Mooresque intuitively given objects are given in whatever experiences it is that we normally can have, not in special experiences called intuitions.

---

[6] *The Theory of Intuition in Husserl's Phenomenology*, André Orianne, trans. (Evanston: Northwestern, 1995; second edition).

But few contemporary philosophers are likely to swallow Moore's "Refutation of Idealism"[7] hook, line, and sinker so as to have the same rationale for their appeals to intuition. Nor are they likely to claim that they have solved the problems to which Moore's methodology led him and his friends when Bertrand Russell and others raised the question as to what the immediately known objects of experience are, for instance, in our sense experience. Sense data? But what precisely are sense data? Surfaces of physical objects? States of one's central nervous system? The search for such immediately given objects in one's experience ended up being truly the longest philosophical journey of the Bloomsbury philosophers.

Thus, contemporary thinkers' practice of appealing to intuitions in philosophical argumentation is without any justification whatsoever—in most cases. One of the interesting exceptions is the main character of the story of twentieth-century intuitions, Chomsky. Unlike his contemporaries, Chomsky could have a good intellectual conscience in appealing to grammatical intuitions, for he is a self-acknowledged Cartesian. He believes in innate ideas, at least in the form of an innate universal grammar. Thus, apart from details, Chomsky had, and presumably has, up his sleeve the same justification for appeals to intuitions as someone like Descartes.

Even though the matter needs closer analysis, it seems that Chomsky's emphasis on intuitions grew in tandem with his keener awareness of the Cartesian character of his own linguistic thinking. Later on, an opposite change seems to have taken place in Chomsky's views in that he has come to put less stress on innateness and universal grammar and, accordingly, less stress on intuitions.

---

[7] *Mind*, N.S. XII (1903): 433-53, reprinted in Moore, *Philosophical Studies* (London. Routledge, 1922), pp. 1-30.

Of course, Chomsky's Cartesianism was originally unknown to the majority of philosophers. He came out of the closet only in his *Cartesian Linguistics*.[8] And when he did, he used the rationalistic line of thought to defend his idea of deep structure rather than his intuitionist methodology. Hence it is perhaps not surprising that the rationale of Chomskian appeals to intuition has not been recognized by philosophers. Philosophers' surprise and dismay at the discovery of Chomsky's rationalism is in evidence in the historical Chomsky-Putnam-Goodman symposium on innate ideas.[9]

Even though Chomsky later enriched his views in a variety of ways and although he does not usually employ the term 'intuition', his position has remained basically the same. For instance, in *Knowledge of Language*,[10] he still postulates "a particular faculty of the mind," the "language faculty." Moreover, speaker's intuitions, now re-named "the judgments of native speakers," provide access to that faculty: "To be sure, the judgments of native speakers will always provide relevant evidence for the study of language" (*ibid.*, p. 37). The main difference is that there are "many different sources apart from judgments concerning the form and meaning of expressions" (*ibid.*, pp. 36-37). But this does not really cast any doubts in Chomsky's mind on competent speakers' intuitions ("judgments") about language. The only qualification Chomsky registers is that such judgments may be disturbed by other factors: "judgments of acceptability, for instance, may fail to produce direct evidence as to grammatical status because of the intrusion of numerous other factors" (*ibid.*, p. 36).

Whether this Cartesian justification of Chomsky's reliance on intuitions is in the last analysis valid or not, it is not available to that great majority of analytic and nonanalytic philosophers who are not Cartesians. Since they do not believe in Aristotelian forms or in Kantian transcendental expositions or deductions as state-of-the-art philosophical truths, their use of intuitions in philosophical argumentation is totally without justification.

Chomsky's reliance on intuitions is a better tactic than philosophers' parallel ploy for two other reasons. First, linguists' intuitions pertain to the human language ability and its products. In contrast, philosophers' intuitions do not pertain to the supposed faculty of intuition itself but to the truths about which this faculty is supposed to provide knowledge. Hence Chomskian intuitions are not subject to the same epistemological problems as philosophers'.

---

[8] New York: Harper, 1966.

[9] *Synthese*, XVII (1967): 2-28.

[10] New York: Praeger, 1986.

Second, Chomsky wisely restricts the intuitions mainly, albeit not exclusively, to one kind only, namely, to intuitions of grammaticality. Admittedly, it turns out, Chomsky notwithstanding, that not even all our intuitions of well-formedness have the kind of generative explanation Chomsky prefers.[11] But linguists are not normally exposed to misdiagnosed intuitions to nearly the same extent as philosophers are.

The example of generative grammarians has been misleading in a more specific respect as well. Typical linguists' intuitions have pertained to the grammaticality of sentences. Linguists have not asked about people's intuitions as to how that grammaticality might be tested. Philosophers' intuitions have typically pertained to the truth or acceptability of various propositions. They have not usually inquired into our intuitions as to how that truth or that acceptability is established. Yet an answer to the latter question is likely to be more easily forthcoming and more confident than an answer to the former. I doubt that many philosophers can honestly claim to have sharp intuitions as to which one of the several so-called theories of truth is itself true, but everyone will agree that in order to ascertain that $(\exists x)S[x]$, one should find an individual $b$ such that $S[b]$; that in order to ascertain that it is the case that $(\forall x)(\exists y)S[x,y]$, one must be able to find, given any individual $d$, an individual $b$ such that $S[d,b]$; and so on.

Philosophers' mistake in adopting the intuitionist interpretation of Chomsky's methodology was compounded by their blind acceptance of an oversimplified construal of another aspect of that methodology. That aspect is the notion of explanation relied on by transformational grammarians. Most analytic philosophers seem to think these days that the main role of their intuitions is to serve as raw material for a generalization. Conversely, a successful generalization—that is, one that captures all the different data provided by intuition—serves to explain those data.

Once again, Chomsky is not a Chomskian, any more than Karl Marx was a Marxist. There is not a trace of the idea of generalization-as-explanation in his early methodological reflections. In his early thinking, paradigm cases of linguistic explanation included accounts of the structural ambiguity of a sentence (like the proverbial 'Visiting relatives can be boring') by pointing that it can be generated in two different ways. In his later writings, he emphasizes such indicators as deductive depth as hallmarks of good explanations at the expense of subsumption under a generalization.

---

[11] See here my "On the Any-thesis and the Methodology of Linguistics," *Linguistics and Philosophy*, IV (1980): 101-22.

It is nevertheless possible to argue that Chomsky's own theories rely far too much on straightforward generalization from examples (particular cases). For instance, the successive versions of the government-and-binding theory seem to have arisen from their predecessors, not through any deeper analysis of the interplay of different factors affecting the conditions of coreference, but merely through adjustments in the relevant generalizations calculated to eliminate counterexamples that had meanwhile been discovered. I have argued that such a tactic reduces greatly the theoretical appeal of generative linguists' theories.[12]

This observation has even some direct methodological significance for linguistics. As I have also argued elsewhere, transformational grammarians and other contemporary linguists would do a much better job if, instead of relying on our intuitions about isolated examples, they tried to vary systematically suitable ingredients in some sample sentence and observed how our "intuitions" change as a consequence. Now we can see why such systematic variation is a way of persuading our intuitions to yield general truths (dependence relations) rather than particular cases.

More generally, reliance on generalization from particular cases is foreign to the methodological spirit of modern science, which originated by looking for dependences of different factors in instructive particular cases (often in controlled experimental situations), and by studying these dependences by the same mathematical means as a mathematician uses in studying the interdependences of different ingredients of geometrical figures in analytic geometry.

In any case, philosophers' reliance on the idea of explanation-by-generalization is rife with historical irony. The best-known philosophical explication of such an idea of explanation—Carl G. Hempel's "covering-law" model of explanation—was generally rejected, at least verbally, around the same time as the intuitionist methodology became popular among them.[13] It has been said that the covering-law model is dead, but in reality it is alive and popular among linguists and among philosophers who are wittingly or unwittingly imitating linguists.

---

[12] See here Gabriel Sandu and my *On the Methodology of Linguistics. A Case Study* (Cambridge: Blackwell, 1991), chapters 4-5.

[13] See here Ilpo Halonen and my "Toward a Theory of the Process of Explanation" (forthcoming).

A closer study of Chomsky's argumentation should, in fact, have discouraged philosophers from generalizing his assumptions beyond linguistic intuitions. In his famous review of B. F. Skinner, Chomsky[14] argues for the necessity of assuming a universal grammar by claiming that without such an assumption, we cannot explain children's capacity for language learning. The argument is that ordinary presuppositionless inductive learning is far too slow to account for the speed of children's language acquisition. But by the same token, it should be necessary for a naturalistic epistemologist to assume strong innate ideas in order to account for the possibility of theory formation by rational scientists. For it is a striking feature of the discovery of many important scientific theories that they relied initially on rather few instances from which they were generalized. An inductive logic that would allow such steps as inductive generalizations presupposes either strong background assumptions or a high degree of a priori faith in the orderliness of nature.

Some naturalistic epistemologists, most prominently W. V. Quine, have acknowledged the need (for their purposes) of some innate principles that could serve to give direction to inductive inferences. For instance, Quine speaks occasionally of innate quality spaces that govern our judgments of similarity. But such Quinean innate ideas are obviously far too few and far apart to provide a basis for a realistic theory of scientific reasoning.

Another dubious ingredient in the Chomskian use of intuitions is the tacit assumption that our basic intuitions about the particular examples to which they pertain are unconscious in the sense that we are not aware of the reasons why we accept or reject them. This feature of Chomskian intuitions is not unrelated to their being restricted to particular cases. The reasons for which we accept and reject particular cases, consciously *or* unconsciously, would realistically speaking have to include some general grammatical rules or principles. But if so, we could become aware of the general rules or principles of grammar directly, without going by way of induction from particular cases, which Chomsky does not allow.

A closer analysis of the nature of scientific explanation confirms these suspicions of contemporary philosophers' methodology. It turns out that, even though there are covering laws hidden in all why-explanations, the actual process of explanation does not consist in subsuming data under such a law. It consists in deriving the explanandum from an established background theory plus suitable ad hoc (or perhaps better ad explanandum) observations. A scientific explanation without a pre-existing theory is an oxymoron.[15] Hempel does receive a qualified vindication, however. It turns out that under a wide range of circumstances, the success of an explanation guarantees the existence of a covering law. But to explain is not to find the covering law, and the covering law is not *the* explanation.

---

[14] Review of *Verbal Behavior, Language*, xxxv (1959): 26-58.

But where do such theories come from? Here we encounter one of the most debilitating weaknesses of contemporary philosophers' reliance on intuitions. For it is generally—though not universally—thought that intuition, like sense perception, always deals with particular cases, however representative. Some such restriction is implicit in all analogies between intuition and sense perception. But if so, intuitions alone cannot yield the general truths: for instance, general theories for which a scientist and a philosopher is presumably searching. Some kind of generalizing process will be needed, be it inductive inference, abduction, or a lucky guess.

This restriction of the data allegedly provided by intuition to particular truths is an instance of the more general assumption that I have called the *atomistic postulate*. According to it, the basic input into our epistemic process consists of particular data, excluding general truths. I have argued that it is the mistake that underlies much of current epistemology and philosophy of science. It is by any judgment an especially debilitating aspect of recent philosophers' intuitionistic methodology. Even if there are intuitions concerning particular cases, we are not any wiser because of them.

A defender of the use of intuitions in philosophy asked me in discussion: "What is wrong with starting from intuitively given obvious particular cases in philosophical argumentation?" I replied: "What would be an example of such obvious truths?" My interlocutor replied: "Surely, everybody agrees, say, that it is wrong to torture a child for pleasure. What is wrong with starting from such universally accepted special cases and using them as a basis for generalizations, in this case ethical ones?"

In fact, I could not have imagined a better illustration of my own point. For what is the generalization implicit in this particular case? This question is tantamount to the question: Which parameters of the given proposition are the ones with respect to which we should generalize? And asking this obviously amounts to asking: What are the factors affecting our judgment about the situation envisaged in it? Is it sometimes all right to torture a child for reasons other than pleasure? If so, what reasons? Is it ever permitted to torture an adult?

---

[15] See footnotes 11 and 13 above.

If so, what is the critical age? Is it ever acceptable to torture an adult for reasons other than pleasure? If so, what reasons? Obviously, my interlocutor's "intuition" does not throw any real light on one's ethical principles.

In order to be useful, intuitions must have some kind of at least implicit generality. For instance, assume that I judge a gift not to be ethically meritorious because it was given merely to satisfy the recipient's expectations. In such a judgment, there is an implicit ethical principle involved in that the judgment presupposes that a necessary condition for an act to be meritorious is that it be spontaneous.

The idea that intuitions pertain to particulars, not to general truths, is deeply entrenched in the history of the idea. It is implied in the traditional consciousness by the analogy between intuition and sense perception; for the majority of philosophers used to follow Aristotle[16] in thinking it is sense perception that deals with particulars. Accordingly, such "intuitive" methods as the use of *ekthesis* in logic were viewed with suspicion. It is interesting to see that thinkers as late as Gottlob Frege rejected all appeals to intuition in logic because intuitions are about particulars, not about the general truths logic deals with.

This misleading idea that intuition deals with particular cases was strongly encouraged by what was perceived as Chomsky's methodology. For the intuitions he recommended linguists to start from were intuitions concerning the grammaticality of particular strings of symbols, not concerning general rules of grammar.

More generally, what is wrong with twentieth-century philosophers' use of their alleged intuitions is not so much that they are wrong as that they are limited in their applicability and that their presuppositions are not recognized and spelled out. In this respect, Kripke's historically influential use of his presumed intuitions offers a telling example. As usual, the problem is not that there is nothing to a philosopher's intuitions. But intuitions in the sense Kripke talks about them come cheaply—at least as cheaply as the *endoxa* from which Aristotle typically starts his argumentation. Indeed, Aristotle's use of *endoxa* offers an interesting object of comparison for contemporary philosophers' handling of their "intuitions." Aristotle realized what Kripke and most others like him have never realized, namely, that *endoxa* or prima facie intuitions do not come fully equipped with instructions for their use. They are not premises for philosophical arguments; they are raw material to be critically weighed, corrected, and integrated into a coherent view. Their presuppositions have to be uncovered and their tacit limitations recognized before such integration is possible. This is what most of Aristotle's own philosophical argumentation amounts to.

---

[16] See *Analytica Posteriora* A 18, 81b6.

Such a critical scrutiny of our prima facie intuitions typically amounts to re-educating our *endoxa* so as to do better justice to the realities of the conceptual situation. One of the most crucial failures of contemporary philosophers has been that they have not realized the need of re-educating and re-regimenting even some of their most sacred intuitions. In this respect, an instructive and clear-cut test case is offered by the intuitions that have prompted the formulation of the most basic laws of our basic logic, the logic of quantifiers. A number of such intuitions were regimented by Frege and Russell into the received logic of quantifiers. It has taken more than one hundred years to realize that that regimentation was too narrow in that it excluded certain perfectly possible and intuitively interpretable structures of quantifiers.[17] The resistance that this simple observation has encountered is unmistakably and undoubtedly due to the fact that it involves a partial re-education of our "intuitions" about quantifiers. In particular, it involves the realization that the full model-theoretical meaning of quantifiers is not exhausted by the "intuitive" idea of their "ranging over" a class of values.

In a different direction, many philosophers still have not acknowledged that the intuitions on which some of the familiar rules of inference are based are not operative in modal and intensional contexts. Logicians' intuitions seemed to tell them, for instance, that existential generalization is a valid mode of inference, that is, that from a sentence of the form $S[b]$ you can always infer $(\exists x)S[x]$. But this 'always' is mistaken. If you confront a philosopher with a specific example, he will readily confess to different intuitions. For instance, the intuitions on which existential generalization is based could never support an inference from:

(1) Ari knows that Homer was Homer.

to:

(2) There is an individual such that Ari knows that he was Homer.

which obviously says the same as:

(2)* Ari knows who Homer was

---

[17] For a fuller explanation, see Sandu and my "A Revolution in Logic?" *Nordic Journal of Philosophical Logic*, 1 (1996): 169-83.

It is equally clear "intuitively" that an inference from (1) to (2)* is valid only if 'Homer' picks out the same individual in all the relevant possible scenarios—in other words, only if Ari already knows who Homer is.

In general, existential generalization with respect to a singular term presupposes that that term refers to the same individual in all the relevant worlds. This is acknowledged, in effect, by the likes of Kripke, who maintains that existential generalization with respect to a proper name is always valid. This he can do because his alleged intuitions include the idea that proper names refer to their objects necessarily, that is, in every possible world. This is what he refers to as rigid designation. In other words, it is not possible that they should refer to one object in one world and to another in another world. In some sense, this intuition can be granted to Kripke. But he never recognizes its tacit limitations, one of which is that it can only be taken to refer to logically or perhaps metaphysically possible worlds. It is the easiest thing in the world to imagine epistemically possible worlds in which a proper name refers to different objects. Thus, to the extent Kripke's basic intuition is applicable to our actual discourse, it excludes from the scope of his theory the most important applications of intensional logics, namely, their application to our epistemic concepts.

What is even more important, Kripke's intuition involves a colossal presupposition. He is assuming that it makes unproblematic sense to speak of the identity of two individuals residing in two different possible worlds. He fails to see that there is nothing in our notion of reference, either intuitively or explicated along the natural possible-worlds lines, that prejudges such questions of cross identity. For the purpose of the reference of, say, a singular term, it suffices to determine the individual (if any) for which it stands in each logically possible world. But such a determination is possible without presupposing anything about which individual is which in another possible scenario. And a modicum of analysis of how our actual discourse operates quickly shows that our principles of cross identification are indeed largely independent of our criteria of reference. We constantly use two different systems of cross identification, which makes nonsense of any unqualified idea of rigid designation.

Kripke's inadequate excuse for not doing the critical analytic work that the explication of his own intuitions requires is that the "possible worlds" with which he is dealing are our constructions out of the individuals that there are in the actual world. But this restricts the applicability of Kripke's intuitions so much as to make them uninteresting. In any half-way realistic application, the different scenarios contain different entities, most of which have no counterparts in the actual world. And these scenarios are not constructed by us in any reasonable (or intuitive) sense of the word. They are typically the objective possibilities we have to heed and from which some nonhuman agent—nature, chance, god—chooses one to be realized. When the National Hurricane Center is preparing for a new season, it assigns proper names to future hurricanes in the order of their appearance, but there is no presumption that in the different possible scenarios for this season, the equally named hurricanes are in some metaphysical sense the same storm. In 1992, it might very well have happened that the first hurricane of the season never touched the mainland of the United States and instead petered out in the Atlantic. I doubt that the good citizens of the Miami area would say that such a possible hurricane was the same as the notorious hurricane Andrew that was the most expensive storm in United States history. Yet it would have had the very same properly conferred name. If Kripke's theory were literally true, the National Weather Service could be accused of grossly abusing the English language by naming hurricanes the way they in fact do. And, what is most important here, if that possible namesake of hurricane Andrew had actually been realized, it would not have been constructed by Kripke as a part of his semantical theory or by any other agent, except (anthropomorphically speaking) by nature.

The upshot is that Kripke's intuitions about what possible scenarios ("worlds") are thus are even more limited in scope than his intuitions about rigid designation. Obviously, in his attempt to visualize his own "relations of ideas" about possible scenarios, Kripke has had in mind only the kinds of artificially simple combinatorial possibilities that probability theorists use in explaining to students their notion of sample space, such as different outcomes of a series of tosses of a coin or a dice or of outcomes of draws of balls from an urn. Unlike Kripke, the probability theorists nevertheless realize fully that most of the real-life applications of their concepts are to sample spaces messier by orders of magnitude than such artificial Las Vegas applications. How unrealistic Kripke is can be shown by the fact that even the simplicity of artificial gambling examples is arguably only skin deep. I have illustrated this point in another (unpublished) paper by reference to the sample space of possible outcomes of a toss of dice:

> [But], what if the high roller in question is King Olaf and his dice splits into two, what do you do? Do you consider both halves of the broken dice as legitimate continuations of the original dice and add up

> what they show, or is only one of them identical with the original one? If so which one? Or is perhaps neither one identical with the old dice?[18] In the first case, an unexpected identificatory situation creates the splitting of sample-space points into several. In the last one, several sample-space points merge with each other. Even small-scale examples [of sample spaces] are thus, in principle, not immune to questions of identification.

What Kripke should have done is to try to work his intuitions patiently and critically into an explicit framework of expressing the intuitions and whatever else he wanted to express. This would have had to include expressing the methods of identification which Kripke presupposed but did not acknowledge. It seems that he might have been prevented from recognizing them by his uncritical adherence to other intuitions. Those intuitions are the ones on which the received Frege-Russell logic of quantifiers are based and which were seen to be unduly restrictive. For what is needed for the purpose of expressing in an obvious way the rigidity of reference by a singular term $b$ in a context governed by the necessity operator $N$? The obvious way is:

(3) $(\exists x)N(b = x)$

For what (3) says is that there is an individual—one and the same individual—that is referred to by $b$ in all possible worlds. But according to the holy rules of the received first-order logic, canonized by the intuitions of most post-Fregean logicians, (1) is entailed by:

(4) $N(b = b)$

which merely says that $b$ is self-identical in each such possible world. Hence an attempted inference from (4) to (3) springs from a confusion between 'each' and 'all'. There is nothing in (4) to guarantee that $b$ is the same individual in different possible worlds. More generally, many philosophers' belief in the universal validity of existential generalization is due to the fact that in "intuiting" this validity, they had before their intuition's eyes a situation involving only what can be said of the actual world (or in any case some one and only possible world). Such intuitions easily become seriously misleading when we are explicitly or tacitly considering several possible scenarios ("worlds"). This nonsensical result makes (3) useless as an expression of rigidity in all logics that assume the validity of existential generalization. Once again, what is needed is a model-theoretical re-education of our intuitions. If (3) expresses what it obviously is calculated to express, namely, that *b* picks one and the same object in all the possible worlds brought into play by the necessity operator *N*, it is not implied by (4). Hence the first thing philosophers discussing identification and reference should do is to reformulate the logic of quantification as it is used in modal and intensional contexts when it comes to existential generalization. This amounts to the recognition of the limited scope of the intuitions which originally prompted the received Frege-Russell logic of quantification, prominently including the rule of existential generalization.

---

[18] Cf. Ivar Ekelund, *The Broken Dice and Other Mathematical Tales of Chance* (Chicago: University Press, 1993), p. 3.

What has been said so far might give the reader the impression that I am against all uses of intuitions in philosophical and scientific reasoning. Such an impression would be wrong. What I am against are the reifications and mystifications surrounding both philosophers' and laymen's ideas about intuitions. These intuitions can be divided into two classes. They are the alleged intuitions concerning empirical truths and those concerning conceptual (including linguistic) ones. (David Hume might have called them intuitions about matters of fact and intuitions about relations of ideas.)

As far as empirical realities are concerned, the brute fact is that the intimations of intuition do not have any privileged epistemological status. They do not carry any automatic justification with them, no matter how convincing they may be subjectively. Epistemologically, they are on the level of clever guesses or perhaps of Aristotle's *endoxa*. We are all familiar (I hope) with insights that are instantly persuasive, but we are also familiar with equal persuasive insights that subsequently turn out to be mistaken.

Furthermore, there is no evidence that such insights, even when valid, are products of any particular mental faculty. Elsewhere, I have shown that what are explicitly called intuitions often turn out to be products of perfectly ordinary discursive thinking, combined with suitable observations. Even when one can locate a characteristic feature of what are called intuitive insights, that special character does not presuppose a separate faculty or capacity. I believe that, especially in mathematical and logical contexts, appeals to intuitions typically amount to uses of model-theoretical insights or concepts. For instance, why is it more "intuitive" to carry out a geometrical proof in terms of figures instead of (say) first-order logic? This question is pertinent because it is perfectly possible to conceive of the use of figures as being nothing but an alternative symbolism parallel with the first-order logical symbolism and intertranslatable with it. A proof in either symbolism can be equally stringent and in fact analogous, so that even in the purely logical proof one can speak, for example, of the introduction of new individuals into the argument in analogy with the "auxiliary constructions" of traditional geometrical proofs. The answer to the question is that the use of figures brings out much more directly the model-theoretical meaning of a geometrical proof.

It is characteristic of the misconceptions that I am criticizing that this "secret" of the intuitiveness of proofs by means of figures has been misinterpreted. It is generally thought that figures served the purpose of facilitating appeals to our geometrical intuition for the purpose of introducing assumptions not sanctioned by the axioms and postulates of geometry. Such appeals are admittedly found in Euclid, but they are as much violations of the rules of figurative argumentation as the corresponding fallacious steps would be violations of the rules of logical proof. Typically, the assumptions so introduced were manifestations of the continuity properties of geometrical objects. Such properties were for a long time difficult for mathematicians and logicians to formulate in any notation.

From these observations it follows that it is nonsense to think that "intuitive" insights must be unconscious. Sometimes they are, but even then it might be possible for the reasoner in question to recognize the discursive steps of reasoning which originally surfaced in the form of a single "intuition." In others, special techniques might be required, ranging from leading questions to truth sera and hypnosis. In others, it is clear that an "intuitive" judgment is the internalized product of training that utilizes perfectly discursive means in teaching the reasoner to notice subtle observational clues. One can enhance the intuitiveness of one's own reasoning quite consciously, as in the use of figures and diagrams. All told, the unavoidably unconscious nature of intuitions is nothing but a pernicious myth we should quickly get rid of.

It may even be the case that we humans are hardwired to carry out certain types of reasoning more "intuitively" than others, probably for evolutionary reasons. For instance, we apparently have fairly sharp intuitions concerning the relationships involved in Euclidean geometry—sharp enough to have led Kant to think that they signal a special status of Euclidean geometry. Such intuitiveness can be exploited for analytic and systematic purposes. For instance, it was a brilliant idea of David Hilbert's and John von Neumann's to think systematically of the relationships among certain functions in analogy with the interrelations of geometrical entities like vectors. But such intuitiveness does not confer any privileged epistemological status on the resulting geometrized theories.

Underlying several of the points just made is an important distinction between heuristics and justification. In a specifiable objective sense, it is easier for human beings to discover and to formulate geometrical proof, for example, by using such "intuitive" techniques as figures. But that does not make such proofs any more or any less stringent than logical proofs conducted in the pedestrian notation of formal logic.

Thus, when it comes to empirical truths and empirical theories independent of my own concepts and my own language, it might very well be wisest to forget the ill-conceived concept of intuition altogether. Of course, this does not deny the reality of the phenomena that people are trying to fit under this term.

But what about my knowledge of my own language and my own concepts? Do I not have, for example, a privileged access to what I mean by my words, of the kind someone might label intuition? This would not be anything stranger than the intuitions of grammaticality on which transformational grammarians at one time used to rely. I have no objection to granting the existence of such access to my own concepts and meanings. What I challenge, however, is the privileged epistemological character of any knowledge obtained through this allegedly privileged access.

How do I find out about another person's language or about her concepts? For instance, what does she mean by a certain word? The first approximation to an answer is: Why not see how that person uses the word? Or perhaps you can simply ask her. If you want to find out how your informant uses color words, put her in a room with differently colored objects and observe which color word she uses for each. Or you can use a color chart and ask the informant which word is applicable to each of them.

This simple (and in some sense simple-minded answer) in fact carries us much further than most philosophers seem to be willing to admit. It may be the case that neither method gives us an indubitable particular case from which to generalize, but I believe that we must adopt a fallibilist epistemology anyway. Quine and some others have argued that we cannot exhaustively determine the meaning of a word in this way. This argument is in the last instance predicated on a rejection of the idea of possible scenario ("possible world"), as is shown in my "Three Dogmas of Quine's Empiricism."[19] If that notion is allowed, then there is, in principle, no definite limit as to how far your experimentation (construction of ever new situations) can carry you in determining the class of scenarios in which the word does or does not apply. And such a determination will, at least for a Montagovian semanticist, determine the meaning of the word. Indeed, in Montague semantics, the meaning of a term is the function that maps possible worlds on references (extensions) of the appropriate logical type (category). And such functions can, in principle, be identified even more and more fully by systematic experimentation with the references that a person assigns to his terms in different actual or imagined scenarios.

---

[19] *Revue Internationale de Philosophie*, LI (1997): 457-77.

The sketchiness of this account is blatant. But this approximate character of my account does not matter here, for I am not trying to give a full account of meaning determination but to compare third-person (or second-person) and first-person cases. And if so, I can simply say that what is semantics for a goose is also for a gander. In other words, the third-person and the first-person cases are, in principle, on a par. I can put myself in different situations and find out how I apply my words in them. Or I can put a question to myself. (There is nothing in the logic of questions and answers that prevents my doing so.) The only difference is that in the first-person cases (of either kind) I can typically skip the stage setting. Instead of putting myself physically in a certain kind of situation and seeing what I shall do in it, I can imagine myself as being in such a situation. Normally, I can confidently say how I would use my words and my concepts in them. Of course, I cannot always do so, as J. L. Austin's[10] bullfinch reminds us, and sometimes I can be wrong. But this does not affect my main point. It is simply an aspect of this analogy between first-person and third-person cases, in that there can be fuzziness in our concepts in either case.

I do not see any reason to deny that I can by means of such a thought experiment obtain objective knowledge about my own language and my own concepts, just as I can do so in the case of others by real experimentation. I do not need a concrete stage setting for the purpose. Thought experiments serve as well in the first-person case as real experiments in the third-person cases. If someone now wants to label the answers obtained by means of such thought experimental "intuitions," I do not have any objections, as long as it is realized what is involved. Such intuitions are not unlike Aristotelian ones, obtained by realizing the appropriate forms (concepts) in my soul. Unlike Aristotelian intuitions, however, the results of my thought experiments need not be infallible. Unlike Aristotle, I am also maintaining that such intuitions are informative only about our own language and about our concepts. Hence there is room in our epistemological arsenal, I believe, for intuitions with a respectable historical pedigree.

---

[10] *Philosophical Papers* (New York: Oxford, 1961), p. 56.

Unfortunately, the vast majority of appeals to intuition by contemporary philosophers cannot be conceived as controlled thought experiments nor be justified by recasting them as such. In view of such goings-on, I am tempted to suggest, half-jokingly—but only *half*-jokingly—that the editors of philosophy journals agree to a moratorium on all papers in which intuitions are appealed to, unless the basis of those appeals is made explicit.

JAAKKO HINTIKKA

Boston University

## COMMENTS AND CRITICISM

# SUPERVALUATION CAN LEAVE TRUTH-VALUE GAPS AFTER ALL*

Among other good things, supervaluation is supposed to allow vague sentences to go without truth values.[1] But Jerry Fodor and Ernest LePore[2] have recently argued that it cannot allow this—not if it also respects certain conceptual truths. The main point I wish to make here is that they are mistaken. Supervaluation can leave truth-value gaps while respecting the conceptual truths they have in mind.

Fodor and LePore are not the only ones to doubt that supervaluation really can do all that it has been held out to do. Supervaluational treatments of vagueness have been said to preserve classical logic, to solve the sorites paradoxes, to account for higher-order vagueness, and more besides. All of this has been called into question.[3] But Fodor and LePore claim to add something new to the case against supervaluational treatments—they claim to have cast doubt on their very *intelligibility*. It will emerge that they have done nothing of the sort.

---

* I would like to thank Steve Kuhn, Alice Kyburg, Ernest LePore, Jeff Pelletier, Achille Varzi, Timothy Williamson, and members of the Logic and Language Group of Georgetown University and the University of Maryland for helpful comments.

[1] Supervaluational treatments of vagueness go back to a suggestion of H. Mehlberg in *The Reach of Science* (Toronto: University Press, 1956). Related ideas can be found in M. Przelecki, "Z Semantyki Pojec Otwartych," *Studia Logica*, XV (1964): 189-220 (translated as "The Semantics of Open Concepts," in J. Pelc, ed., *Semiotics in Poland, 1896-1969* (Dordrecht: Reidel, 1979), pp. 284-317); Michael Dummett, "Wang's Paradox," *Synthese*, XXX (1975): 301-24; David Lewis, "General Semantics," in Donald Davidson and Gilbert Harman, eds., *Semantics of Natural Language* (Boston: Reidel, 1972, 2nd ed.), pp. 169-218; and Hans Kamp, "Two Theories about Adjectives," in E.L. Keenan, ed., *Formal Semantics of Natural Language* (New York: Cambridge, 1975), pp. 123-55. The supervaluational approach is taken furthest by Kit Fine, in "Vagueness, Truth and Logic," *Synthese*, XXX (1975): 265-300.

[2] "What Cannot be Evaluated Cannot be Evaluated and It Cannot Be Supervalued Either," this JOURNAL, XVIII, 10 (October 1996): 516-35.

[3] Timothy Williamson argues that supervaluations do not preserve classical logic—*Vagueness* (New York: Routledge, 1994). Kamp criticizes the supervaluational solution to the sorites paradoxes—"The Paradox of the Heap," in U. Mönnich, ed., *Aspects of Philosophical Logic* (Boston: Reidel, 1981), pp. 157-74; so does R.M. Sainsbury—*Paradoxes* (New York: Cambridge, 1988). Criticisms of the supervaluational treatment of higher-order vagueness can be found in Williamson; and in Sainsbury, "Concepts without Boundaries," in R. Keefe and P. Smith, eds., *Vagueness: A Reader* (Cambridge: MIT, 1996), pp. 251-64.

0022-362X/99/9603/148–56

I shall proceed as follows. In section I, I set out supervaluational interpretations of vague languages; in section II, I state Fodor and LePore's objection to these; finally, in section III, I argue that this objection is groundless. Not only *can* supervaluation leave truth-value gaps while respecting Fodor and LePore's conceptual truths; we shall see that it *must* leave them, if it is to respect these truths.

## I SUPERVALUATIONAL TREATMENTS OF VAGUENESS

To begin with, I shall expound the supervaluation of vague sentences. I shall follow Kit Fine's treatment (*op. cit.*). This discussion is a supplement to Fodor and LePore's, which oversimplifies at one crucial point that I shall get to shortly.

Underlying supervaluational treatments is the idea that vagueness is a variety of semantic incompleteness. Take the vague adjective 'bald'. It is semantically incomplete insofar as it can happen that, besides those individuals which are determinately bald and those which are determinately not bald, some borderline cases are neither. Coupling a vague adjective with the name of a borderline case results in a sentence—say, 'Al is bald'—that is in turn semantically incomplete: it is neither true nor false; it has a *truth-value gap*.

Supervaluational treatments depend furthermore on the idea that vagueness can be reduced by resolving indeterminacy. With a vague adjective that happens when, leaving determinate cases untouched, borderline cases are resolved one way or the other. For example, 'bald' can be made more precise in various ways, accordingly as borderline cases are included among the determinately bald or among the determinately not-bald. The different ways of making vague expressions more precise are *precisifications*. With a view to making things concrete, we can regard them as assignments of suitable extensions to vague expressions; so each precisification assigns to 'bald' a positive extension of individuals that are determinately bald, and a negative extension of individuals that are determinately not-bald. One precisification $p^*$ is *more precise* than another, $p$, if the positive extensions that $p^*$ assigns to expressions are more inclusive than the positive extensions that $p$ assigns to them, while the negative extensions that $p^*$ assigns are also more inclusive than the corresponding negative extensions that $p$ assigns.

The intuitively understood meanings of vague expressions limit the ways of making them more precise. *Internal* constraints control the boundaries of single predicates: if on making 'bald' more precise we count one man in, then we must also count in every man with even less hair. Other *external* constraints control the common boundaries of several predicates: making 'red' and 'pink' more precise shall not lead us to count the same blob as both. Fine calls these two kinds of constraints *penumbral connections*. Precisifications within the limits of penumbral connections are *admissible*.

Fine models precisifications after the disambiguations of ambiguous expressions. He seems to think that vagueness and ambiguity are closely related:

> [T]he only formal difference is that the precisifications may be infinite, even indefinite, and may be subject to penumbral connection. Vagueness is ambiguity on a vast and systematic scale (*op. cit.*, p. 282).

The idea that vagueness is on a par with ambiguity is surely controversial, though. For one thing, whereas disambiguations are *contextual*, insofar as we typically assert just one disambiguation of an ambiguous sentence, precisifications are not contextual in quite this way. We do not pick just one complete precisification of a vague sentence and assert that.

Even so, precisifications might be somewhat contextual. Consider the possibilities for reference with vague definite descriptions. Suppose Al is at a bar being salient along with Mick Jagger, David Bowie, and others who, unlike Al, still have a lot of hair. Then someone spots him and points him out: "Look, there's Al! He's the bald one." Now, Al is not *really* bald. He is borderline. Even so, this utterance of 'the bald one' denotes him. This can be accounted for if we suppose that, although we never pick one complete precisification, we sometimes do pick a more or less limited range of them. We can suppose that this utterance leads us to pick those ways of drawing a precise boundary which include, among the bald, a single salient individual; for, in keeping with the relevant penumbral connection, each such precisification will single out Al. Our picking these precisifications we can regard as an accommodation, furnished to enable the definite description to denote.

Precisifications, if somewhat contextual, are in this regard at least somewhat like disambiguations. Be this as it may, there is no need to tie vagueness to ambiguity here, and some advantage to keeping them separate. We can simply think of the precisifications of a vague expression as all of the different ways of drawing precise boundaries that resolve borderline cases. The question whether vagueness and ambiguity are related can be left to one side.

So much for precisifications. I shall now turn to the evaluation of vague sentences. On a supervaluational treatment, the truth of, say, 'Al is bald' depends on two things. One is Al's hair. The other is the precisification of 'bald'. Let us put aside questions of temporality and modality, so that we need not consider the hair Al once had but no longer has, or hair he might have had but actually does not. Then we can think of the nonlinguistic matters on which truth depends, such as Al's head of hair, as fixed. Truth will depend only on the different precisifications of vague expressions.

Supervaluation amounts, as Fine puts it, to "a sort of principle of non-pedantry: truth is secured if it does not turn upon what one means" (*op. cit.*, p. 278). The idea is that a vague sentence shall be counted true if it is true on each way of making it more precise. On the simplest view, the only precisifications that matter are the *complete* ones, which leave nothing at all indeterminate. It is there that evaluation begins. Evaluation at a complete precisification is classical. Each individual falls either within the positive extension of any given predicate, or within its negative extension, so each atomic sentence can be assigned a determinate truth value, *true* or *false*. Sentences compounded with the truth-functional connectives 'and', 'not', and so on follow suit, according to the classical truth definitions. In this way, each complete precisification determines a classical valuation.

Moving now from complete precisifications to those which retain some indeterminacy, truth goes by the principle of nonpedantry. A sentence is counted true if it is true on each way of making it completely precise:

> *Super-truth condition*: $S$ is true (/false) on precisification $p$ just in case for every admissible complete $p^*$ that is more precise than $p$, $S$ is true (/false) on $p^*$.

A sentence will go without a truth value on $p$ if it is true on some admissible complete precisifications of $p$, but false on others. So an incomplete precisification determines a *partial* valuation, on which some sentences are true, others are false, and yet others are neither.

We evaluate sentences relative to a preferred precisification that is *appropriate*, insofar as it retains a measure of indeterminacy that reflects the intuitively understood meanings of vague expressions. Fodor and LePore call such an appropriate precisification a *base model*, or simply a *model* (518-19). But the super-truth condition calls on other, complete precisifications. An appropriate precisification and these others together make up an *interpretation*, or what Fine calls a *specification space*. An interpretation is a tree with the appropriate precisification at the base of the trunk, with increasingly precise ones as we move upward through the branches, and with complete precisifications at the leaves. Truth, falsity, and indeterminacy go by the appropriate precisification: a sentence is true in an interpretation if true on each of its complete precisifications; false if false on each of them; otherwise, the sentence is indeterminate.

It is now possible to say at which point Fodor and LePore oversimplify. They interpret vague language in models that are, in effect, isolated precisifications. My interpretations and Fine's specification spaces, on the other hand, include a range of different precisifications. In section III, it will become clear that Fodor and LePore's treatment does not lend itself to a proper understanding of the expression 'there is no matter of fact about whether'. Meanwhile, in the following section, we shall see that this matter lies right at the heart of their objection.

## II. FODOR AND LEPORE'S OBJECTION

It would appear that supervaluation will leave 'Al is bald' without a truth value, if Al is borderline bald. Not so, say Fodor and LePore; not if supervaluation honors the principle:

(P) Conceptual truths must be respected by all classical models, including classical valuations (521).

Fodor and LePore argue that there are conceptual truths which (P) demands classical valuations respect but which, if respected, rule out complete precisifications. This will in turn rule out truth-value gaps, since for a sentence to go without a truth value there have to be complete precisifications on which it is true, and others on which it is false.

To give an example, we must first decide just what kind of thing Fodor and LePore take conceptual truths to be: perhaps sentences, or the contents of sentences, or something else again? They do not say.[4] Plausibly conceptual truths are sentences;[5] in this case, the following is a representative example:

(1) There is no fact of the matter about whether a man one ninth of whose head is covered with hair is bald (522).

Nothing depends on the fraction one ninth, of course. You can substitute any other you like, if you think a man with that much hair is a borderline case. Fodor and LePore argue that we have here not merely a truth, but indeed a *conceptual* truth. For the sake of the argument, let us allow that it is so.

---

[4] What Fodor and LePore do say does not settle this: "[W]e use 'conceptual truths' in (P) as a dummy term for contributions that nonlogical vocabulary makes to meaning" (521). Sentences, maybe; their contents, maybe. But *contributions?*

[5] What, otherwise, are we to make of (P)? It is to sentences that valuations assign truth values, not to their contents, or to anything else.

Fodor and LePore's conceptual truths must be respected by all classical valuations, according to (P). So they must hold true on all complete precisifications (since these determine classical valuations). That is, what (P) demands is that these truths shall enjoy the favored status of Fine's *penumbral truths*, whose truth on all complete precisifications is secured by penumbral connections. An example is: 'If one man is bald, then so is every other man with even less hair'. Fodor and LePore do not offer a generally compelling reason to accept this demand of (P). They do not really need to, though, since it is sufficient for their purpose that the *supervaluationist* must accept it; and this, it seems, he must. Honoring (P) seems to be the motivation for introducing penumbral constraints in the first place, and if the supervaluationist is out to honor this principle at all, he ought to honor it fully. He cannot go picking and choosing between conceptual truths, expecting some of them to hold on all complete precisifications, but not others.

But fully honoring (P) is the undoing of supervaluational treatments of vagueness, according to Fodor and LePore. They argue that it trivializes such treatments by ruling out complete precisifications. Indeed, it might well appear that they are onto something. To ensure that Fine's penumbral truths are respected, we rule out all precisifications on which they fail to hold true. This is the effect of introducing penumbral constraints. But if we try to ensure respect for Fodor and LePore's conceptual truths in the same way, it might seem, we will end up with no complete precisifications at all. Take (1) for example. How, one might wonder, can this sentence hold true on a complete precisification? Complete precisifications resolve all indeterminacy; they place every individual within either the positive or the negative extension of every predicate. And where there is no indeterminacy, there is a fact of the matter about *everything*. So, one might think, on a complete precisification there is a fact of the matter about whether a man, one ninth of whose head is covered with hair, is bald. Since this is the very thing that (1) denies, it might seem that (1) cannot hold true on a complete precisification.

## III HOW TO HAVE TRUTH-VALUE GAPS AND ALSO RESPECT CONCEPTUAL TRUTHS

This reasoning should be resisted. Precisifications do resolve indeterminacy, and complete precisifications resolve it completely. But to resolve indeterminacy is not to remove it, since where indeterminacy can be resolved in a range of different ways, it remains present in the multiplicity of different ways of resolving it. This idea is the basis of my criticism of Fodor and LePore's argument.

Fodor and LePore's conceptual truths assert that there is a certain indeterminacy. I argue that what is required, if they are to be respected, is variety among precisifications. For example, respecting (1) requires that there be two kinds of complete precisifications: some on which men one ninth of whose heads are covered with hair are included among the bald, and others on which such men are not included. On a proper understanding, I argue, respecting Fodor and LePore's conceptual truths does not require ruling out precisifications, in addition to those ruled out by Fine's penumbral constraints. On the contrary, what it requires is ruling precisifications *in*, to safeguard variety. This completely undercuts Fodor and LePore's objection. As we shall see, honoring (P) is not merely compatible with truth-value gaps. It guarantees that there are some.

Let us turn first to the sentential operator 'it is indeterminate whether'. Fodor and LePore use it interchangeably with 'there is no matter of fact about whether'; I shall soon use it to render their conceptual truths. Consider a sentence with this operator, say, 'It is indeterminate whether Al is bald'. This sentence is to hold true if Al is borderline bald, which is to say if some complete precisifications include Al among the bald, while others do not.

Notice that this is what might be called a *global* truth condition. An interpretation, comprising a space of precisifications, is the sort of thing that can satisfy it. An isolated precisification is not. One might think that, having global truth conditions, sentences with the operator 'it is indeterminate whether' cannot be evaluated on the individual precisifications of an interpretation. In this case, these sentences will somehow have to be evaluated without the super-truth condition. Maybe Fodor and LePore's conceptual truths, rendered with this operator, can be treated in such a way: evaluated, indeed validated, but not supervalued.

In fact, though, it is straightforward to evaluate the relevant sentences on individual precisifications. According to the following truth condition, a precisification that satisfies such a sentence does so in virtue of its place within a space of suitably various precisifications:

> *Truth condition for 'it is indeterminate whether'*: 'It is indeterminate whether S' is true on a complete precisification, in any given interpretation, if S is true on some complete precisifications of the interpretation, but false on others.

This truth condition ignores the higher-order vagueness that arises where borderlines themselves have borderlines. For example, 'bald' is vague in the second-, third-, and higher degrees if one might be a borderline case of a borderline bald man, a borderline case of a borderline case of a borderline bald man, and so on. To take higher-order vagueness into account, we could allow that 'It is indeterminate whether Al is bald' can be true on some ways of making 'bald' more precise (those in which, once all higher-order vagueness has been resolved, Al lands in the borderline of 'bald') but false on others (in which he lands elsewhere). Clearly, my truth condition does not allow this. Higher-order vagueness does not seem to be relevant to this discussion, though, and I shall now put it aside. My criticism of Fodor and LePore's argument can also be made, at greater length, while taking it into account.

Given the restriction to first-order vagueness, I claim, this is the proper treatment of 'it is indeterminate whether'.[4] Notice that Fodor and LePore's version of supervaluation does not lend itself to this treatment. Once this operator is taken into account, truth on a complete precisification is a global matter, involving a space of precisifications. But, as I remarked in section I, Fodor and LePore seem to think that, on a supervaluational approach, vague language is to be interpreted in models that are, in effect, isolated precisifications. At this crucial point they oversimplify.

Returning now to Fodor and LePore's conceptual truths, we can take (1) as a representative example and render it as follows:

> For each *x*, if *x* is a man and one ninth of *x*'s head is covered with hair, then it is indeterminate whether *x* is bald.

This illustrates my claim that, as seen from a supervaluational perspective, Fodor and LePore's conceptual truths hold true in virtue of the variety to be found among precisifications. Sentence (1) holds true (on one complete precisification of an interpretation, indeed on all such precisifications and, by the super-truth condition, in the interpretation itself) provided there are two kinds of complete precisifications: some in which men one ninth of whose heads are covered with hair fall within the positive extension of 'bald', and others in which they fall within the negative extension. This conceptual truth is respected, in the sense of principle (P), provided all interpretations of English include precisifications of these two kinds.

---

[4] Others agree. Dummett (*op. cit.*, p. 311) and Fine (*op. cit*, p. 289) both suggest a treatment of the sentential operator 'it is definitely the case that' which parallels the possible-worlds treatment of necessity: letting 'D' stand for this operator, a sentence *DS* will true (on any given precisification) just in case S holds true on all relevant precisifications. Fine (*op. cit.*, pp. 287-89) goes on to define in terms of *D* an operator *I* that is equivalent to 'it is indeterminate whether'. He gives an ingenious treatment allowing for higher-order vagueness (*op. cit.*, pp. 293-94, but watch for typographical errors). Williamson repeats this treatment (*op. cit.*, p. 159).

The problem with Fodor and LePore's objection is now in plain view. They claim that honoring (P) rules out complete precisifications, and with them truth-value gaps. But it is simply not so. Interpretations of English can respect their conceptual truths while having truth-value gaps. An interpretation that includes complete precisifications of the relevant two kinds respects (1), and leaves 'Al is bald' without a truth value: this sentence is true on precisifications of the first kind, in which Al falls within the positive extension of 'bald'; it is false on those of the second kind, in which he falls within the negative extension; so, by the super-truth condition, this sentence goes without a truth value. In fact, honoring (P) is not merely compatible with truth-value gaps. It *guarantees* that there are some, since if (1) is to hold true on each complete precisification of each interpretation of English, then *each* such interpretation will have to have complete precisifications of the relevant two kinds.[7] There are sure to be truth-value gaps, given some named men with the right amount of hair.

So Fodor and LePore are quite mistaken. Among other good things, supervaluation can leave truth-value gaps while validating the conceptual truths they have in mind.

MICHAEL MORREAU

University of Maryland/College Park

---

[7] Strictly speaking, each *nontrivial* interpretation, having complete precisifications, will have to have them.

## NEW BOOKS: ANTHOLOGIES

EVERSON, STEPHEN, ed.: *Ethics: Companions to Ancient Thought: Volume 4.* New York: Cambridge, 1998. 322 p. Cloth $64.95, paper $22.95.

EZE, EMMANUEL CHUKWUDI, ed.: *African Philosophy: An Anthology.* Malden, MA: Blackwell, 1998. xiii, 494 p. Paper $29.95.

FIRTH, RODERICK: *In Defense of Radical Empiricism: Essays and Lectures by Roderick Firth.* JOHN TROYER, ed. Lanham, MD: Rowman and Littlefield, 1998. 439 p. Cloth $70.00, paper $26.95.

FISCHER, JOHN MARTIN, and MARK RAVIZZA, S.J., eds.: *Responsibility and Control: A Theory of Moral Responsibility.* New York: Cambridge, 1998. viii, 277 p. Cloth $59.95.

FLETCHER, ANDREW: *Political Works.* JOHN ROBERTSON, ed. New York: Cambridge, 1997. xlix, 243 p. Cloth $64.95, paper $24.95.

FLEW, ANTONY: *Philosophical Essays.* JOHN SHOSKY, ed. Lanham, MD: Rowman and Littlefield, 1998. 213 p. Cloth $58.00, paper $22.95.

FREELAND, CYNTHIA A., ed.: *Feminist Interpretations of Aristotle.* University Park: Penn State UP, 1998. xiv, 369 p. Cloth $55.00, paper $19.95.

FRENCH, STANLEY G., WANDA TEAYS, and LAURA M. PURDY, eds.: *Violence against Women: Philosophical Perspectives.* Ithaca: Cornell, 1998. 260 p. Cloth $49.00, paper $16.95.

GARBER, DANIEL, and MICHAEL AYERS, eds.: *The Cambridge History of Seventeenth Century Philosophy: Volumes I and II.* New York: Cambridge, 1998. xvii, 1616 p. Cloth $175.00.

GARRETT, DON, and EDWARD BARBANELL, eds.: *Encyclopedia of Empiricism.* Westport, CT: Greenwood, 1997. xvii, 455 p. Cloth $99.50.

GAUKROGER, STEPHEN, ed.: *The Soft Underbelly of Reason: The Passions in the Seventeenth Century.* New York: Routledge, 1998. vi, 167 p. Cloth $65.00.

GENTZLER, JYL, ed.: *Methods in Ancient Philosophy.* New York: Oxford, 1998. viii, 398 p. Cloth $72.00.

GOOD, JAMES, and IRVING VELODY, eds.: *The Politics of Postmodernity.* New York: Cambridge, 1998. ix, 232 p. Cloth $59.95, paper $18.95.

GRIFFIN, STEPHEN M., and ROBERT C.L. MOFFAT, eds.: *Radical Critiques of the Law.* Lawrence: Kansas UP, 1997. x, 337 p. Cloth $45.00, paper $25.00.

GUPTA, CHHAND, and D.P. CHATTOPADHYAYA, eds.: *Cultural Otherness and Beyond.* Boston: Brill, 1998. ix, 252 p. Cloth $94.50.

GUTMANN, AMY, ed.: *Freedom of Association.* Princeton: University Press, 1998. vi, 382 p. Cloth $65.00, paper $19.95.

HABER, JORAM G., and MARK S. HALFRON, eds.: *Norms And Values: Essays on the Work of Virginia Held.* Lanham, MD: Rowman and Littlefield, 1998. vii, 296 p. Cloth $62.50, paper $23.95.

HAHN, LEWIS EDWIN, ed.: *The Philosophy of Roderick M. Chisholm.* Chicago: Open Court, 1997. xviii, 738 p. Paper $36.95; *The Philosophy of P.F.*

*Strawson.* Chicago: Open Court, 1998. xviii, 428 p. Cloth $72.95, paper $42.95.

HANNAY, ALASTAIR, and GORDON D. MARINO, eds.: *The Cambridge Companion to Kierkegaard.* New York: Cambridge, 1997. xi, 428 p. Cloth $59.95, paper $18.95.

HECK, RICHARD, JR., ed.: *Language, Thought, and Logic: Essays in Honor of Michael Dummett.* New York: Oxford, 1997. vii, 308 p. Cloth $77.00.

HICKMAN, LARRY A., ed.: *Reading Dewey: Interpretations for a Postmodern Generation.* Bloomington: Indiana UP, 1998. xxi, 271 p. Cloth $39.95, paper $17.95.

HJORT, METTE, and SUE LAVER, eds.: *Emotion and the Arts.* New York: Oxford, 1997. 302 p. Cloth $49.95, paper $19.95.

HOBBES, THOMAS: *On The Citizen.* RICHARD TUCK and MICHAEL SILVERTHORNE, eds. New York, Cambridge, 1998. lii, 250 p. Cloth $59.95, paper $18.95.

HULL, DAVID L., and MICHAEL RUSE, eds. *The Philosophy of Biology.* New York: Oxford, 1998. ix, 772 p. Cloth $85.00.

JACKSON, FRANK, ed.: *Consciousness.* Cambridge: Harvard, 1998. xii, 506 p. Cloth $55.00.

JADACKI, JACEK JULIUSZ, and WITOLD STRAWIŃSKI, eds.: *In the World of Signs: Essays in Honour of Professor Jerzy Pelc.* Atlanta: Rodopi, 1998. 459 p. Cloth $120.50, paper $23.50.

JAGGAR, ALISON M., and IRIS MARION YOUNG, eds.: *A Companion to Feminist Philosophy.* Malden, MA: Blackwell, 1998. xvii, 703 p. Cloth $84.95.

JENKINS, KEITH, ed.: *The Postmodern History Reader.* New York: Routledge, 1998. xiii, 443 p. Paper $24.99.

JUNG, CARL GUSTAV: *Jung's Seminar on Nietzsche's* Zarathustra. JAMES L. JARRETT, ed. Princeton: University Press, 1998. xxviii, 393 p. Paper $17.95.

JUTRONIĆ, DUNJA, ed.: *The Maribor Papers in Naturalized Semantics.* Maribor: Slovenia UP, 1997. 428 p.

KEEFE, ROSANNA, and PETER SMITH, eds.: *Vagueness: A Reader.* Cambridge: MIT, 1997. 352 p. Cloth $35.00.

KELLY, PAUL, ed.: *Impartiality, Neutrality and Justice: Re-Reading Brian Barry's* Justice as Impartiality. Edinburgh: University Press, 1998. xi, 261 p. Cloth $27.50.

KNELLER, JANE, and SIDNEY AXINN, eds.: *Autonomy and Community. Readings in Contemporary Kantian Social Philosophy,* Albany: SUNY, 1998. xi, 334 p. Paper $21.95.

KOERTGE, NORETTA, eds.: *A House Built on Sand: Exposing Postmodernist Myths about Science.* New York: Oxford, 1998. xi, 322 p. Cloth $29.95.

KOURANY, JANET A., ed.: *Philosophy in a Feminist Voice: Critiques and Reconstructions.* Princeton: University Press, 1998. xii, 322 p. Cloth $50.00, paper $16.95.

KUHSE, HELGA, and PETER SINGER, eds.: *A Companion To Bioethics.* Malden, MA: Blackwell, 1998. xv, 512 p. Cloth $99.95.

LAMB, ROGER, ed.: *Love Analyzed.* Boulder: Westview, 1996. xvii, 267 p. Cloth $58.00, paper $17.95.

LANDES, JOAN B., ed.: *Feminism: The Public and the Private.* New York: Oxford, 1998. x, 507 p. Cloth $75.00, paper $19.95.

LAWRENCE, CHRISTOPHER, and STEVEN SHAPIN, eds.: *Science Incarnate: Historical Embodiments of Natural Knowledge.* Chicago: University Press, 1998. vii, 342 p. Paper $19.00.

LENZER, GERTRUDE, ed.: *Auguste Comte and Positivism: The Essential Writings.* New Brunswick, NJ: Transaction, 1998. lxxxii, 555 p. Paper $29.95.

LÉON, CÉLINE, and SYLVIA WALSH, eds.: *Feminist Interpretations of Søren Kierkegaard.* University Park: Penn State UP, 1997. xviii, 347 p. Cloth $32.50, paper $17.95.

LEVIN, DAVID MICHAEL, ed.: *Language beyond Postmodernism: Saying and Thinking in Gendlin's Philosophy.* Evanston, IL: Northwestern, 1997. x, 378 p. Cloth $89.95, paper $29.95.

LOCKE, JOHN: *Locke: Political Essays.* MARK GOLDIE, ed. New York: Cambridge, 1997. xl, 409 p. Cloth $59.95, paper $19.95.

LOTT, TOMMY L., ed.: *Subjugation and Bondage: Critical Essays on Slavery and Social Philosophy.* Lanham, MD: Rowman and Littlefield, 1998. xxii, 349 p. Cloth $57.00, paper $22.95.

LUDLOW, PETER, and NORAH MARTIN, eds.: *Externalism and Self-Knowledge.* Stanford: CSLI, 1998. x, 382 p. Cloth $69.95, paper $27.95.

LYONS, DAVID, ed.: *Mill's Utilitarianism: Critical Essays.* Lanham, MD: Rowman and Littlefield, 1997. xiv, 193 p. Cloth $34.00, paper $12.95.

MACHAMER, PETER, ed.: *The Cambridge Companion to Galileo.* New York: Cambridge, 1998. xii, 462 p. Cloth $59.95, paper $19.95.

MARCUSE, HERBERT: *Technology, War, and Fascism.* DOUGLAS KELLNER, ed. New York: Routledge, 1998. xvi, 278 p. Cloth $35.00.

MASON, ANDREW, ed.: *Ideals of Equality.* Malden, MA: Blackwell, 1998. xi, 114 p. Paper $24.95.

MAY, LARRY, CHRISTINE SISTARE, and JONATHAN SCHONSHECK, eds.: *Liberty, Equality, and Plurality.* Lawrence: Kansas UP, 1997. Cloth $45.00.

MCKEON, RICHARD: *Selected Writings of Richard McKeon: Volume I: Philosophy, Science, and Culture.* ZAHAVA K. MCKEON and WILLIAM G. SWENSON, eds. Chicago: University Press, 1998. 528 p. Cloth $50.00.

MCKIM, ROBERT, and JEFF MCMAHAN, eds.: *The Morality of Nationalism.* New York: Oxford, 1997. xii, 371 p. Cloth $45.00, paper $19.95.

MCNEILL, WILLIAM, and KAREN S. FELDMAN, eds.: *Continental Philosophy: An Anthology.* Malden, MA: Blackwell, 1998. viii, 458 p. Paper $34.95.

MILL, HARRIET TAYLOR: *The Complete Works of Harriet Taylor Mill.* JO ELLEN JACOBS and PAULA HARMS PAYNE, eds. Bloomington: Indiana UP, 1998. xliv, 604 p. Cloth $59.95.

MILLER, DANIEL, ed.: *Material Cultures: Why Some Things Matter.* Chicago: University Press, 1998. xi, 243 p. Cloth $45.00, paper $22.00.

MITCHAM, CARL, ed.: *Research in Philosophy and Technology: Volume XVI: Technology and Social Action.* Greenwich, CT: JAI, 1997. xv, 285 p. Cloth $78.50.

MONTAG, WARREN, and TED STOLZE, eds.: *The New Spinoza.* Minneapolis: Minnesota UP, 1997. xx, 260 p. Cloth $62.95, paper $24.95.

MORAVCSIK, JULIUS M., ed.: *Meaning, Creativity, and the Partial Inscrutability of The Human Mind.* Stanford: CSLI, 1998. viii, 200 p. Cloth $64.95, paper: $23.95.

MORRIS, CHRISTOPHER W., and JULES L. COLEMAN, eds.: *Rational Commitment and Social Justice: Essays for Gregory Kavka.* New York: Cambridge, 1998. viii, 252 p. Cloth $54.95.

MOSS, JEREMY, ed.: *The Later Foucault.* Thousand Oaks, CA: Sage, 1998. x, 201 p. Paper $22.95.

MURASUGI, KUMIKO, and ROBERT STAINTON, eds.: *Philosophy and Linguistics.* Boulder: Westview, 1998. ix, 285 p. Cloth $65.00.

NUYTS, JAN, ed.: *Language and Conceptualization.* New York: Cambridge, 1997. 281 p. Cloth $59.95.

NYE, ANDREA, ed.: *Philosophy of Language: The Big Questions.* Malden, MA: Blackwell, 1998. xii, 305 p. Cloth $62.95, paper $32.95.

OSTERTAG, GARY, ed.: *Definite Descriptions: A Reader.* Cambridge: MIT, 1998. xii, 411 p. Paper $30.00.

PARKER, IAN, ed.: *Social Constructionism, Discourse and Realism.* Thousand Oaks, CA: Sage, 1998. xii, 159 p. Paper $22.95.

PATTISON, GEORGE, and STEVEN SHAKESPEARE, eds.: *Kierkegaard: The Self in Society.* New York: St. Martin's, 1998. xii, 225 p. Cloth $55.00.

PAUL, ELLEN FRANKEL, FRED D. MILLER, JR., and JEFFREY PAUL, eds.: *Virtue and Vice.* New York: Cambridge, 1998. xiii, 326 p. Paper $21.95.

PEIRCE, CHARLES S.: *The Essential Peirce: Selected Philosophical Writings, Volume 2: 1893-1913.* THE PEIRCE EDITION PROJECT, eds. Bloomington: Indiana UP, 1998. xxxviii, 584 p. Cloth $47.00, paper $19.95; *Chance, Love, and Logic: Philosophical Essays.* MORRIS R. COHEN, ed. Lincoln: Nebraska UP, 1998. xlv, 318 p. Paper $15.00.

PHILLIPS, D.Z., ed.: *Rush Rhees on Religion and Philosophy.* New York: Cambridge, 1997. xxii, 389 p. Cloth $69.95.

PLANTINGA, ALVIN: *The Analytic Theist: An Alvin Plantinga Reader.* JAMES F. SENNETT, ed. Grand Rapids, MI: Wm. B. Eerdmans, 1998. xviii, 369 p. Paper $30.00.

RACHELS, JAMES, ed.: *Ethical Theory.* New York: Oxford, 1998. vii, 578 p. Cloth $39.95, paper $19.95.

RAZAVI, MEHDI AMIN, and DAVID AMBUEL, eds.: *Philosophy, Religion, and the Question of Intolerance.* Albany: SUNY, 1997. xvii, 284 p. Paper $19.95.

REID, THOMAS: *An Inquiry into the Human Mind on the Principles of Common Sense: A Critical Edition.* DEREK R. BROOKES, ed. University Park: Penn State UP, 1997. xxv, 345 p. Cloth $55.00.

ROGERS, KELLY, ed.: *Self-Interest: An Anthology of Philosophical Perspectives.* New York: Routledge, 1997. 293 p. Cloth $75.00, paper $24.99.

ROSENFELD, MICHEL, and ANDREW ARATO, eds.: *Habermas on Law and Democracy: Critical Exchanges.* Berkeley: California UP, 1998. xii, 466 p. Cloth $55.00.

ROYCE, JOSIAH: *Metaphysics: Josiah Royce: His Philosophy 9 Course of 1915-1916, as Stenographically Recorded by Ralph W. Brown and Complemented by Notes from Bryon F. Underwood.* WILLIAM ERNEST HOCKING, RICHARD HOCKING, and FRANK OPPENHEIM, eds. Albany: SUNY, 1998. xxi, 346 p. Paper $24.95.

SAUNDERS, JASON L., ed.: *Greek and Roman Philosophy after Aristotle.* New York: Free Press, 1997. 371 p. Paper $18.95.

SCHOTT, ROBIN MAY, ed.: *Feminist Interpretations of Immanuel Kant.* University Park: Penn State UP, 1997. xiii, 423 p. Cloth $55.00, paper $18.95.

SCHRIFT, ALAN D., ed.: *The Logic of the Gift: Toward an Ethic of Generosity.* New York: Routledge, 1997. xii, 334 p. Cloth $74.95, paper $19.95.

SHAPIRO, IAN, and ROBERT ADAMS, eds.: *Integrity and Conscience: Nomos Volume XL.* New York: University Press, 1998. xii, 340 p. Cloth $50.00.

SHRADER-FRECHETTE, KRISTIN, and LAURA WESTRA, eds.: *Technology and Values.* Lanham, MD: Rowman and Littlefield, 1997. xi, 472 p. Cloth $69.50, paper $26.95.

SILVERMAN, HIGH J., ed.: *Cultural Semiosis: Tracing the Signifier.* New York: Routledge, 1997. 325 p. Cloth $75.00, paper $22.99.

SLOOP, JOHN M., and JAMES P. MCDANIEL, eds.: *Judgment Calls: Rhetoric, Politics, and Indeterminacy.* Boulder, CO: Westview, 1998. vii, 263 p. Cloth $59.00.

STERBA, JAMES P., ed.: *Ethics: The Big Questions.* Malden, MA: Blackwell, 1998. vii, 455 p. Cloth $62.95, paper $32.95.

STEWART, JON, ed.: *The Debate between Sartre and Merleau-Ponty.* Evanston, IL: Northwestern UP, 1998. xlvii, 634 p. Cloth $89.95, paper $24.95.

STEWART, M.A., ed.: *Studies in Seventeenth-Century European Philosophy.* New York: Oxford, 1997. ix, 237 p. Cloth $65.00.

STOCK, GUY, ed.: *Appearance versus Reality: New Essays on the Philosophy of F.H. Bradley.* New York: Oxford, 1998. x, 236 p. Cloth $65.00.

TAYLOR, C.C.W., ed.: *Oxford Studies in Ancient Philosophy, Volume XV.* New York: Oxford, 1997. 294 p. Cloth $68.00.

TAYLOR, MARK C., ed.: *Critical Terms for Religious Studies.* Chicago: University Press, 1998. 423 p. Paper $18.00.

TURNER, SUSAN M., and GARETH B. MATTHEWS, eds.: *The Philosopher's Child: Critical Essays in the Western Tradition.* Rochester: University Press, 1998. x, 238 p. Cloth $65.00.

VAN OPHUIJSEN, JOHANNES M., and MARLEINE VAN RAALTE, eds.: *Theophrastus: Reappraising the Sources, Rutgers University Studies in Classical Humanities, Volume VIII.* New Brunswick, NJ: Transaction, 1998. ix, 410 p. Cloth $54.95.

VINCENT, ANDREW, ed.: *Political Theory: Tradition and Diversity.* New York: Cambridge, 1997. xi, 275 p. Cloth $59.95, paper $22.95.

VOEGELIN, ERIC: *History of Political Ideas, Volume II: The Middle Ages to Aquinas.* PETER VON SIVERS, ed. Columbia: Missouri UP, 1997. xxvi, 248 p. Cloth $34.95.

WEITHMAN, PAUL, ed.: *Religion and Contemporary Liberalism.* Notre Dame: University Press, 1997. viii, 315 p. Cloth $48.00, paper $23.00.

WILLET, CYNTHIA, ed.: *Theorizing Multiculturalism: A Guide to the Current Debate.* Malden, MA: Blackwell, 1998. xi, 466 p. Paper $26.95.

WILLIAM, OF OCKHAM: *William Ockham: Opera Politica IV.* H.S. OFFLER, ed. New York: Oxford, 1997. xviii, 486 p. Cloth $150.00.

WRIGHT, CRISPIN, BARRY C. SMITH, and CYNTHIA MACDONALD, eds.: *Knowing Our Own Mind.* New York: Oxford, 1998. x, 450 p. Cloth $65.00.

ZACK, NAOMI, LAURIE SHRAG, and CRISPIN SARTWELL, eds.: *Race, Class, Gender, and Sexuality: The Big Questions.* Malden, MA: Blackwell, 1998. xiii, 410 p. Cloth $62.95, paper $32.95.

## NOTES AND NEWS

The Editors regret the untimely death of the American philosopher Nelson Goodman on November 25, 1998, at the age of 92. He received his B.S. in 1928 and his Ph.D. in 1941, both from Harvard University. Goodman was a Professor of Philosophy at Harvard University from 1968-1977. Previously he served as Professor of Philosophy at Brandeis University from 1964-1967, Professor from 1951-1964 and Associate Professor from 1946-1951 at the University of Pennsylvania, and Instructor at Tufts College, 1944-1945. His books include *A Study of Qualities* (1941), *The Structure of Appearance* (1951), *Fact, Fiction, and Forecast* (1955), *Languages of Art* (1968), *Problems and Projects* (1972), *Ways of Worldmaking* (1978), *Of Mind and Other Matters* (1984), *Reconceptions in Philosophy and Other Arts and Sciences*, with Catherine Z. Elgin (1988), *L'Art en théorie et en action* (1996). Throughout his lifetime, Goodman was the recipient of many honors and awards. He served as President of the American Philosophical Association, Eastern Division, in 1967, and as Vice President of the Association for Symbolic Logic, 1950-1952. He was a Fellow of the American Academy of Arts and Sciences and a corresponding Fellow of the British Academy. Among his many distinguished lectures and lecture series were: Alfred North Whitehead Lecture, Harvard University, 1962; John Locke Lectures, Oxford University, 1962; Miller Lectures, University of Illinois, 1974; Immanuel Kant Lectures, Stanford University, 1976; and the Howison Lecture, University of California, Berkeley, 1985. Goodman's activities related to art extended beyond philosophy. He was the founder, and director of Project Zero at Harvard, 1967-1971, an interdisciplinary research program investigating aesthetic education. He was also producer of the Arts Orientation Series, 1969-1971, and Director of the Dance Center, both at Harvard. The Philosophy Department at Harvard is reportedly planning a colloquium in his honor, possibly in the spring or fall of 1999.

THE JOURNAL OF PHILOSOPHY
VOLUME XCVI, NO. 4, APRIL 1999

UNIVERSITY LIBRARY PUR CAMPUS

# IS TWO A PROPERTY?*

"Whatever is, is one...whatever are, are many."
—Bertrand Russell (1903)[1]

Bill and Hillary live together in Washington. Bill is one human being. So is Hillary. But they are two humans; they are different humans. What is it for them to be two humans? It is for them to instantiate a property, namely, being two humans. This is the view that I aim to clarify, defend, and develop in this paper. In doing so, I reject the standard conception of property (and relation) and defend a broader conception that allows nonstandard properties like the one in question. The standard conception retains the long-standing bias against the many in favor of the one; the conception contains principles to the effect that, roughly, many things as such cannot instantiate a property. The broader conception results from rejecting those principles.

I begin by clarifying the notion of the many in section I. In section II, I formulate and argue against the crucial principles of the standard conception. In section III, I clarify the language used to attribute nonstandard properties to the many. Then I elaborate on the broader conception, with the aid of the resources available in the language, in section IV. Section V is an epilogue that places the conception in a historical perspective.

---

* I am grateful to Joseph Almog, Tyler Burge, Kit Fine, David Kaplan, Bernard Linsky, Penelope Maddy, Tony Martin, and Graham Priest for their help at various stages of writing this paper. My work on this paper was partially supported by a Killam Postdoctoral Fellowship at the University of Alberta and a University of Queensland Postdoctoral Research Fellowship, which are hereby gratefully acknowledged.

[1] *Principles of Mathematics* (New York: Norton, 1937, 2nd edition), p. 132.

0022-362X/99/9604/163–90

## I. MANY VERSUS ONE

Here is the contrast between what is one and what are many. Bill is *one*. He is some one thing, specifically, one human being. Likewise with Hillary. But Bill and Hillary are *many* things, namely, more than one thing. They are, specifically, *two* humans. Moreover, the two humans are many things only, not some one thing. By contrast, Bill is one thing only, not many things. He (or his body) *has* many parts: one heart, two eyes, four limbs, and the like. But this does not make Bill himself the many parts, any more than his having two parents makes him the two parents. Surely, the contrast does not rest on the fact that Bill and Hillary are both humans. Anything, any object, is one thing. It is irrelevant to this what kind of thing it is, whether it is an adult or a child, a human or a city, material or immaterial, spatiotemporal or not, concrete or abstract, and so on. Likewise, Bill and Hillary are two simply because they are not identical with each other. Whether they are similar or not, how closely related they are, and so on are immaterial to their being many.

Some might object that the contrast somewhat breaks down because there are *complex objects*, such as sets or classes, groups or collections, aggregates or mereological sums, or the so-called totalities or pluralities. They agree that such an object, too, is one thing, but object that it is many things as well. This is not correct. Consider, for example, the set {Bill, Hillary}, an object that in some sense comprehends both Bill and Hillary (but nothing else). The set is two, the objectors argue, because it *is* (namely, is identical with) the two humans (though it is not either of them). But this assumption does not hold. Bill and Hillary are two humans; by contrast, the set is not two humans.[2] The set in a sense *has* two humans; it is related to them as its members. But this does not make the set itself two humans; it does so no more than Bill's having two Americans as parents makes Bill himself two Americans.[3] Thus a complex object, too, is one thing only, not many things. How many members, elements, or parts it has is as irrelevant to this as is how many parents a human has.

---

[2] One might object that one cannot get '{Bill, Hillary} *is* two humans' from 'Bill and Hillary *are* two humans' by replacing 'Bill and Hillary' with '{Bill, Hillary}'. Although I think the sentences have the same predicate 'to be two humans', which takes different *forms*, we can bypass this issue by considering a contrast immune to the objection: *Bill and Hillary*, and Chelsea are three humans; *{Bill, Hillary}* and Chelsea are not three humans (they are only two things: one set and one human)

[3] In *Parts of Classes* (Cambridge: Blackwell, 1991), David Lewis holds that the mereological sum of Bill and Hillary is in a sense identical with the two humans (so it is in a sense two humans). I have argued against this view in "Is Mereology Ontologically Innocent?" *Philosophical Studies*, XCIII, 2 (February 1999): 141-60.

Some might respond by replacing, in effect, the talk of being two (or being many) with the talk of related properties that complex objects instantiate, such as *being two-membered*, a property instantiated by any set (or class or whatever has members) with exactly two members. One cannot identify this latter property with the former, they concede; being two-membered is instantiated by the set {Bill, Hillary}, which it is not correct to say instantiates being two. Still, they argue, one can hold that being two-membered is the real predicative component of facts like the following:

(1) Bill and Hillary are two.

That is, on their view, the sentence 'Bill and Hillary are two' states the same fact that the sentence '{Bill, Hillary} is two-membered' states, that is, the fact that:

(2) {Bill, Hillary} is two-membered.

but the former sentence is misleading about the nature of the fact because there is no property indicated by the apparent predicate 'to be two'.[4]

This view rests on an incorrect analysis of facts like (1). To see this, consider facts like the following:

(3) Bill is not identical with Hillary [and Bill is a thing and Hillary is a thing].

Bill and Hillary must be two different things, *insofar as* Bill is not identical with Hillary (while Bill exists as a thing and so does Hillary). So:

Thesis I: Fact (1) is a logical consequence of fact (3).

But just because Bill is not identical with Hillary (while both Bill and Hillary exist) does not mean:

---

[4] For an account of (natural) number based on this analysis, see Penelope Maddy, *Realism in Mathematics* (New York: Oxford, 1990). For example, she identifies the number two as the property of being two-membered. Fregean accounts that identify numbers with a kind of objects (for example, $\{\emptyset, \{\emptyset\}\}$) can also be seen to be based on analyzing facts like (1) as facts like (2) that contain *relational properties* characterized with reference to the objects (for example, being equinumerous with $\{\emptyset, \{\emptyset\}\}$). For Frege's account, which differs from the contemporary Fregean accounts in identifying numbers as extensions of concepts, see, for example, his *Die Grundlagen der Arithmetik* (Breslau: Köbner, 1884), translated by J. L. Austin as *The Foundations of Arithmetic* (New York: Blackwell, 1950). In *A System of Logic* (New York: Harper, 1891, 8th edition), J. S. Mill identifies, for example, two as a property (having two parts) instantiated by an aggregate. This account, too, must rest on a similar analysis of facts like (1).

(4) There is something (for example, a set) of which Bill is a member and Hillary is a member.

Thus, fact (4) is *not* a logical consequence of fact (3).[5] Because, however, fact (4) *is* a logical consequence of fact (2):

Thesis II: Fact (2) is *not* a logical consequence of fact (3).

Now, it follows from theses I and II that fact (1) cannot be reduced to fact (2).

Underlying this argument is the thought that fact (1) is a fact which concerns both Bill and Hillary but which concerns nothing else; Bill and Hillary are two (consequently, many) things simply because they are things not identical with each other. The fact that Bill is identical with himself does not concern the singleton {Bill}, which is not Bill, because the fact concerns nothing but Bill. Likewise, fact (1) does not concern the set {Bill, Hillary}, which is neither Bill nor Hillary, no matter how closely it might be related to both Bill and Hillary. It is the same with any complex object, be it the aggregate of the humans or the so-called plurality thereof. Thus, fact (1) cannot be analyzed as involving a complex object.

The proper analysis of facts like (1), I shall argue, is the one implicit in the above distinction between the one and the many. Fact (1) is composed of (a) *Bill and Hillary*, on the one hand, and (b) *being two*, on the other; its predicative component, being two, is a property instantiated by many things (for example, Bill and Hillary) but not by any one of them. Similarly, the fact that Bill and Hillary are many is formed by being many, a property instantiated by no object, simple or complex. Accepting this analysis requires a radical departure from the standard conception of property and relation.

## II BEING TWO AS A PLURAL PROPERTY

On the standard conception of property and relation, there is no property like being two. I shall now explain two related principles in the conception which rule out such properties, and argue that facts like (1) cannot be given a proper analysis without rejecting the principles.

---

[5] One might object to this by identifying logical consequence with modal entailment, defined via necessity (namely, *metaphysical* necessity, as it is commonly called). On the orthodox view on the modal status of sets, fact (3) modally entails fact (4); see, for example, Kit Fine, "First-order Modal Theories: Sets," *Noûs*, XV (1981): 177-205. This, however, gives no relief for the analysis in question. The fact that Bill is a human and the fact that {Bill} *has exactly one member that is a human* modally entail each other, on the orthodox view, but the two facts cannot be identified, because the property of *having exactly one member that is a human* cannot be taken to be the predicative component of the former fact. Thus, it is necessary to draw a distinction among facts finer than the one available through modal entailment. My view is that logical consequence, among others, yields such a distinction. For more on this or related issues, see Fine, "Essence and Modality," *Philosophical Perspectives*, VIII (1994): 1-16; and Joseph Almog, "The What and the How," this JOURNAL, LXXXVIII, 5 (May 1991): 225-44.

*II.1. The principle of singular instantiation.* On the standard conception, a property cannot be instantiated by many things *as such*; there is no such property as being two, which would be instantiated by two things *as such*. For a better understanding of reasons for rejecting such properties, it is necessary to be clearer about the content of the phrase 'as such'. The phrase is necessary because the conception accommodates some properties that are in a sense instantiated by many things: for example, *being a human*, which is in a sense instantiated by, for example, Bill and Hillary. What is the difference between the standard properties like this and those ruled out by the conception?

One might attempt to explain it as follows. Being a human, instantiated (in a sense) by Bill and Hillary, is instantiated by every one of them; being two, by contrast, is not instantiated by every one of those, by whom it is supposed to be instantiated. But this explanation does not get at the heart of the matter. *Being one or more humans* would be instantiated by any one of those who instantiate it; either of Bill and Hillary, who are one or more humans (specifically, two humans), is one or more humans (specifically, one human). But the standard conception rules out such a property as well. This is due to an important feature which it shares with, for example, being two and distinguishes them from the standard properties.

Being a human must *occur more than once* to be instantiated by Bill and Hillary—whether to form two different facts, such as the fact that Bill is a human and the fact that Hillary is a human, or as different components of one complex fact, such as the fact that:

> Bill *is a human* [the first occurrence], and Hillary *is a human* [the second occurrence].

By contrast, being two can (or could) occur just once to be instantiated by many things; it should be taken to occur only once to be instantiated by the two humans and, thus, form fact (1). It is the same with being one or more humans; this property can (or could) occur just once to be instantiated by Bill and Hillary—though it can also occur two more times to be instantiated by Bill and Hillary, respectively.

Thus, the standard conception contains the following principle:

> The principle of singular instantiation: there is no property that can occur just once to be instantiated by many things (that is, more than one thing).

*II.2. The principle of singularity.* The principle of singular instantiation is, I think, a consequence of an underlying principle that formulates the more general constraint that the standard conception imposes on relations as well as properties. To state this principle, it is useful to clarify some basics about objects, properties, and relations.

Bill, for example, is an object; being a human, a property; *living in*, a relation. There is a categorical difference between objects and properties or relations. Very roughly, properties or relations are *predicable*; objects are not. Being a human is predicated of Bill (specifically, Bill *instantiates* the property); Bill *is a human*. Similarly, living in is predicated of Bill and Washington, so to speak, in the given order (specifically, Bill *bears* the relation *to* Washington); Bill *lives in* Washington. Bill, by contrast, cannot be predicated at all. He occurs in the fact that Bill is Bill, but not as a predicable; he occurs in it to bear the relation of identity, indicated[6] by 'is' in the sentence 'Bill is Bill', to himself.[7]

Properties are distinguished from relations by the number of *argument places*. A property has one argument place; a relation, two or more. Being a human has only one argument place; it can be filled in with Bill to yield the fact that Bill is a human. Living in has two argument places; they can be filled in with Bill (the first) and Washington (the second), respectively, to yield the fact that Bill lives in Washington.

The distinction between properties and relations, we can see, is not as fundamental as the one between objects and properties or relations. The difference between a property and a two-place relation is no greater than the one between the relation and a three-place relation, such as the one indicated by the predicate 'to give...to'. So I shall henceforth use the term 'relation' in a broader sense to apply it to properties as well: properties are one-place relations, relations with one argument place.

Now, consider what can fill in argument places of relations. The argument place of being a human can be filled in with (in short, admits of) Bill because he instantiates the property. In general, whatever instantiates a property can fill in its argument place. But the converse does not hold. The argument place of a property admits of whatever instantiates its complement as well, because a property and its complement have the same kind of argument places. London, for example, can fill in the argument place of being a human, whose complement, the property of not being a human, is instantiated by the city.

---

[6] I use 'indicate' widely for the relation between any linguistic expression and its metaphysical counterpart, and reserve 'refer' for the more specific relation applicable to nonpredicable expressions.

[7] There is also the property of *being Bill* (composed of identity and Bill), but Bill differs from the property.

If so, what cannot fill in the argument place of a relation? The argument place of being a human does not admit of many things *as such*. It (in a sense) admits of Bill and Hillary, but for this to happen, the property must occur twice; its argument place admits of only one thing (that is, Bill) *as the property occurs just once*. And it is the same with relations like living in. The first argument place of this relation (in a sense) admits of Bill and Hillary to yield the fact that:

> Bill and Hillary *live in* Washington.

But the relation must occur twice for this to happen, as can be seen in the following analysis of the fact:

> Bill *lives in* Washington, and Hillary *lives in* Washington.

Now, on the standard conception, this is the general situation with any argument place of any relation whatsoever; that is:

> The principle of singularity: an argument place of a relation does not admit of many things *as the relation occurs just once.*

We can see that this explains why, on the conception, the principle of singular instantiation holds; a property cannot occur just once to be instantiated by many things unless its argument place admits of them as the property occurs just once.

*II.3. The ground for plural relations.* I shall now argue that the principle of singularity does not hold. Given the argument, it is straightforward to see that the principle of singular instantiation does not hold, either.

Call an argument place of a relation *plural* if it admits of many objects as such; *singular* otherwise. Accordingly, call a relation *singular* if all of its argument places are singular; *plural* otherwise. On the principle of singularity, there is no plural relation. I argue that there are plural relations because there are facts like the following:

> Bill and Hillary *are two*. (Fact (1))
> Bill and Hillary *are two humans.*
> Bill and Hillary *live together.*
> Russell and Whitehead *coauthor* a book.
> Some humans *love only one another.*
> The coauthors of *Principia Mathematica* (in short, PM ) *cooperate.*

Bill *is* not *one of* the coauthors of PM.

The proper analysis of such facts, I claim, involve plural properties like:

being two
being two humans
living together
cooperating
loving only one another

and two-place relations like:

coauthoring
being one of

whose first and second argument places, respectively, are plural.

My argument for the claim runs as follows. (A) Being two, for example, is the predicative component of fact (1) (so it is a relation), indicated by the predicate 'to be two', that it shares with facts like the following:

(5a) Dolly and Washington *are two.*
(5b) Washington and Neptune *are two.*

(B) And it has only one argument place (so it is a property), which admits of Bill and Hillary to yield fact (1). (C) Finally, the argument place is a plural one because the property occurs only once in, for example, fact (1), in which the argument place is filled in with many things, for example, Bill and Hillary.

One might essentially agree with (A) but object that the predicative component of fact (1) is the singular two-place relation indicated by the phrase '... and —— are two'.[8] This view might seem plausible if one considers only facts like (1) or (5a), but it fails to square with facts like the following:

(6a) Bill, Hillary, and Washington *are not two.*
(6b) Bill, Hillary, Washington, and Neptune *are not two.*

These facts cannot be seen to contain a singular two-place relation. Their common predicative component, the relation of *not being two,* must be taken to have only one argument place that admits of Bill, Hillary, and Washington (as such) to yield fact (6a); and Bill, Hillary, Washington, and Neptune (as such) to yield fact (6b). That is, not being two is a plural property. Now, because this is the complement of the predicative component of fact (1), being two, this must also be a plural property.

---

[8] For an account of number that rests on this view, see John Bigelow, *The Reality of Numbers* (New York: Oxford, 1988), pp. 49-54. (I have added 'essentially' because he might object that the phrase, strictly speaking, does not indicate being two.)

Some might concede that being two is not a two-place relation, one with two argument places *built into* it, but object that it is a singular relation of a special kind, one that Henry Leonard and Nelson Goodman[9] call a *multigrade relation*. On this view, being two (or its complement) is a relation that can occur in one fact with, for example, two argument places filled in while occurring in another fact with another number of argument places filled in. But this view meets difficulty with facts like the following:

(7) There are some things that *are not two*.

which shares the complement of being two with facts like (6a). How many, and what kind of, argument places does the relation occur with in fact (7)? The right answer to this question is: it occurs with only one argument place, a plural one. Then there is no reason to take the relation to occur with other, singular argument places in, for example, fact (6a).

How else can one object to the above argument for plural properties? It is hard to reject (C) while granting (A)-(B). One cannot argue that being two occurs twice in fact (1), which cannot be analyzed as a conjunctive fact. Thus, to reject (C), one would have to argue that Bill and Hillary are in a sense one complex object, but this, we have seen, is not a correct view. Some might object that there is no fact corresponding to the construction 'Bill and Hillary are two', because this is not grammatically correct. For, on their view, 'two' used as an adjective must be complemented with nouns (or noun phrases). But this view does not help much in defending the principle of singularity. One can apply the above argument to, for example, the fact that:

(8) Bill and Hillary are two humans.

and conclude that being two humans is a plural property. Because they cannot challenge this argument with the grammatical objection, they might attempt to do so by defending alternative analyses of fact (8). But the analyses they might propose can be seen to result from different ways of analyzing fact (1), because fact (8) can be analyzed as containing fact (1) as a conjunct as follows:

---

[9] See their "The Calculus of Individuals and Its Uses," *Journal of Symbolic Logic*, v, 2 (June 1940): 45-55. The linguistic counterparts of multigrade relations are called *variably polyadic predicates*. For more discussions of the views examined in this paragraph and the previous, see my "Numbers and Relations," *Erkenntnis*, XLIX, 1 (1998): 93-113

Bill and Hillary *are two*, and they are humans.

Can one then defend a way to analyze facts like (1) or (8) without plural relations? One might hold that such facts contain only singular properties instantiated by complex objects. On her view, fact (1) can be analyzed as follows:

(2) {Bill, Hillary} is two-membered.

which results from the set {Bill, Hillary} filling in the argument place of the singular property of being two-membered. But such an analysis cannot be right, we have seen, because fact (1) concerns nothing (that is, no object) but Bill or Hillary.

To avoid this problem, one might attempt to analyze such facts in terms of properties rather than complex objects. They, one might argue, are second-order facts like the one composed of (i) the singular (first-order) property of *being identical with Bill or Hillary* and (ii) a second-order property that somehow corresponds to 'two', one that is instantiated by any singular property instantiated by two things.[10] For one can argue (correctly, on my view) that this fact does not logically entail the existence of any *object* except Bill or Hillary; the existence of a *property*, which it logically entails, does not logically entail the existence of an object that one might use as its representative.

Because properties are not determined by their instances, however, the scheme of second-order analysis meets difficulties in preserving logical relations among facts like the following:

(9) Russell and Whitehead cooperate.
(10) Russell and Whitehead are the authors of PM
(11) The authors of PM cooperate.

Consider the second-order facts that the scheme renders them as:

(9a) *cooperate* (is identical with Russell or Whitehead)[11]
(10a) $\forall\alpha[\alpha$ is identical with Russell or Whitehead $\leftrightarrow \alpha$ is an author of PM]
(11a) *cooperate* (is an author of PM)

---

[10] This is the kind of analysis that would be proposed by, for example, Michael Dummett, who argues that "a plural noun-phrase...under a correct analysis...is seen to figure predicatively"—*Frege: Philosophy of Mathematics* (Cambridge: Harvard, 1991), p. 93.

[11] The sentence used here is a second-order predication whose predicate is '*cooperate*', whose argument place is filled in with the predicate 'is identical with Russell or Whitehead'.

Although the properties indicated by the phrases 'is identical with Russell or Whitehead' and 'is an author of PM' are co-extensional (so, fact (10a) holds), they are not the same property; to be identical with Russell or Whitehead is not the same as to be an author of PM. If so, one of the properties may instantiate a second-order property—for example, the one indicated by '*cooperate*'—that the other does not. Thus, (11a) is not a logical consequence of (9a) and (10a).[12] But, clearly, (11) is a logical consequence of (9) and (10).[13]

To avoid this problem, one might analyze facts like (9) not as second-order predications but as second-order existential generalizations like:

(9b) $\exists X[\forall\alpha(X(\alpha) \leftrightarrow \alpha \text{ is identical with Russell or Whitehead}) \wedge \textit{cooperate}(X)]$[14]

But this modified scheme fails other logical relations. Consider the following:

(12) Russell and Whitehead are philosophers who coauthor PM.
(13) There are some philosophers that cooperate and coauthor PM.

The scheme renders fact (12) as follows:

(12a) $\exists X[\forall\alpha(X(\alpha) \leftrightarrow \alpha \text{ is identical with Russell or Whitehead}) \wedge [\forall\alpha(X(\alpha) \rightarrow \alpha \text{ is a philosopher}) \wedge \textit{coauthor}(X, \text{PM})]]$

This fact and fact (9b) do not logically entail the second-order counterpart of fact (13):

---

[12] One cannot defend the scheme by adopting the extensional conception of relation (which is implausible given properties like the above-mentioned) that identifies co-extensional relations. (11a) is not a logical consequence of (9a) and (10a) unless the "identity" between the first-order properties is a logical consequence of their extensional equivalence. But it is one thing to hold the extensional conception; quite another to hold (more implausibly) that its truth relies on logic alone.

[13] One might analyze (9) as the fact that Russell cooperates-with Whitehead, and Whitehead cooperates-with Russell. This analysis is not applicable to fact (11), which has the same predicative component. One can avoid this problem by analyzing facts like (9) and (11) in terms of the second property indicated by '$\forall\alpha[..\ \alpha\ .. \rightarrow \exists\beta(\alpha \neq \beta \wedge \alpha \text{ cooperates-with } \beta)]$', whose argument place (indicated by '...$\alpha$...') admits of a first-order property (for example, being identical with Russell or Whitehead). But this analysis fails because pairwise cooperation of, for example, three humans does not logically entail their cooperation: any two of them may cooperate (for example, to write a book between them while excluding the other, or to kill the other) while not all three of them cooperate.

[14] This is the key idea in Russell's "No-Class Theory," on which the talk of sets (or "classes") is reducible to the talk of properties (or "propositional functions"); see, for example, his *Introduction to Mathematical Philosophy* (London: Allen and Unwin, 1919), p 187f.

(13a) $\exists X[\forall\alpha(X(\alpha) \rightarrow \alpha \text{ is a philosopher}) \wedge [\textit{cooperate}\ (X) \wedge \textit{coauthor}\ (X, \text{PM})]]$

But, clearly, facts (9) and (12) logically entail fact (13).[15]

Now, some people might concede that fact (9) contains a plural property (and thus, reject the principle of singularity) but still object that fact (1) should be analyzed as a second-order fact whose predicative component is a special kind of second-order property, the one that corresponds to the quantifier phrase '$\exists\alpha\exists\beta(\alpha\neq\beta \wedge \forall\gamma[\ldots\gamma\ldots \leftrightarrow (\gamma=\alpha \vee \gamma=\beta)])$'. On their view, the phrase 'Russell and Whitehead' functions ambiguously in 'Russell and Whitehead are two' and 'Russell and Whitehead cooperate'; and the predicative components of the facts stated by them cannot combine to yield a complex property, because they do not have the same kind of argument places. But there is nothing wrong with contracting the two sentences as if the phrase functions unambiguously therein, and the resulting sentence states a logical consequence of the two facts:

Russell and Whitehead are two and cooperate.

which logically entails the fact that:

There are some things that are two and cooperate.

Being two, like cooperating, must be seen to occur as a plural property in this fact—as a component of the complex property: *being two and cooperating*. Then facts like (1), too, must be analyzed in terms of the plural property given their connection to the more complex facts like the above.

This completes my argument against the principle of singularity. It should be clear from the argument that the principle of singular instantiation must also be rejected. Not every property that violates the former principle violates the latter principle; being one has an argument place that admits of, for example, Bill and Hillary as such, but is not instantiated by them as such. But the properties mentioned in the argument are counterexamples to the latter principle as well. Be-

---

[15] One might consider analyzing, for example, fact (9) as follows:

$$\exists X(\forall\alpha[X(\alpha) \leftrightarrow \alpha \text{ is identical with Russell or Whitehead}] \wedge \forall Y[\forall\alpha(X(\alpha) \leftrightarrow Y(\alpha)) \rightarrow \textit{cooperate}(Y)])$$

This relies on the assumption that the property indicated by '*cooperate*' is instantiated by any property co-extensional with a property that instantiates it. But one cannot assume this without analyzing the second-order property. For more on second-order treatments, see subsection III 2.C below.

ing two, for example, is instantiated by any two things as such; its complement, by any things more than two.

Notice also that this last point establishes that the argument place of being two admits of any two or more things as such—no matter how many they are, whether they are finitely many or infinitely so.[16] Moreover, the argument place, like that of being one, admits of any one thing—for example, Bill—as well. For any one thing instantiates the property of not being two because there are facts like the following:

> Bill *is not two* [in particular, not two humans].
> Bill *is not two*, but only one [namely, one human].

## III THE LANGUAGE AND LOGIC OF PLURALS[17]

In the next section, I shall elaborate on the nature of plural relations and confirm the intuitions adduced against alternative analyses of facts like (1), such as the intuition that fact (1) is a logical consequence of fact (3). I prepare for the task here by clarifying the syntax, semantics, and logic of the language I shall use for that purpose. The language contains plural, as well as singular, constructions, which allows one to use it to refer to many things as such and indicate and analyze plural, as well as singular, relations. We are familiar with the language. It is English (or its relevant fragment).

*III.1. Informal syntax and semantics.* English has: *singular terms*, such as the proper name 'Bill', the definite description 'the author of *Academica*', or the pronoun 'it'; *singular predicates*, predicates all of whose argument places are *singular* (namely, they admit of only a singular term), such as 'to be *a* human'; and *singular quantifiers*, such as 'something' (or 'There is something that'). In addition, it has: *plural terms*, such as the conjunctive plural term 'Bill and Hillary', the plural definite description 'the authors of PM', or the plural pronoun 'they'; *plural predicates*, predicates some of whose argument places are *plural* (that is, they admit of a plural term[18]), such as one-place predicates like 'to be two humans' or 'to cooperate', or two-place predicates like 'to be one of', 'to write', or 'to carry'; and *plural quantifiers*, such as 'some thing*s*' (or 'There are some things that'). Thus, the language contains sentences like 'Bill and Hillary are two humans', 'Russell is one of the authors of PM', and 'There are some things that are two humans, and they cooperate and write a book', as well as sentences like 'Bill is a human', 'Cicero is the author of *Academica*', and 'There is something that is a human, and he writes a book'.

---

[16] Infinitely many things (for example, the natural numbers) cannot (as such) be indicated by a plural term resulting from joining their names using 'and', but the absence of such a term no more makes the things unsuitable for the plural property than the absence of a term referring to a pebble makes the pebble unsuitable for a singular property.

[17] For elaborations of discussions in this section, see my "The Language and Logic of Plurals," *Synthese* (forthcoming).

[18] Plural argument places may admit of a singular term as well; recall, for example, 'to be two'. (One might distinguish them further into *neutral* and *exclusively plural* ones.) Most English predicates, on my view, have such an argument place; 'Cicero writes *Academica*' and 'Russell and Whitehead write PM' have the same predicate

Plural terms must be sharply distinguished from (first-order) predicates. These are *predicable* expressions that can combine with nonpredicable expressions to form a sentence. The predicate 'to be a human' combines with the singular term 'Bill' to yield the sentence 'Bill is a human'. By contrast, plural terms, like singular terms, are nonpredicable expressions. The plural term 'Bill and Hillary' cannot combine with 'Bill' to yield a sentence, just as the singular term 'Bill' cannot combine with 'Bill' to yield a sentence. Nonpredicable expressions can constitute a sentence—for example, 'Bill is one of Bill and Hillary' or 'Bill is Bill'—only by combining with predicable expressions—for example, 'to be one of' or 'to be'.

Accordingly, different kinds of semantic relations are appropriate for plural terms, on the one hand, and predicates, on the other. The relation of *reference* is appropriate for nonpredicable expressions; *being true of*, for predicables (of one argument place).[19] The predicate 'to be a human' (or 'to be one of Bill and Hillary') is true of Bill, but does not refer to him; the singular term 'Bill', by contrast, refers to Bill, but is not true of him. How about plural terms? They are referential expressions. The plural term 'Bill and Hillary', for example, is not true of anything (or any things). It refers to some things, namely, Bill and Hillary (as such).[20] This does not mean that the plural term refers to Bill and also to Hillary; it refers to neither of them. A typical[21] plural term refers to some things without referring to any one of them.

We can characterize the truth of plural, as well as singular, predications in terms of the semantic functions of terms and predicates.

---

[19] There is another semantic function of predicables: they indicate relations. We need to consider this function to characterize the truth of sentences with, for example, second-order predicates (except first-order quantifiers), but the fragment of English here in question contains no such predicate.

[20] 'To refer to' as used here is a plural predicate, which indicates a plural relation; likewise with 'to be true of' as used below to be applicable to plural predicates.

[21] This qualification is necessary because some plural terms refer to only one thing 'Cicero and Tully' is a plural term, but refers to only one thing: Cicero, that is, Cicero and Tully, who are not many.

The singular predication 'Bill is a human' is true if its predicate is true of Bill, to whom the term 'Bill' refers. To give a similar account of plural predications like 'Bill and Hillary are two', we need to consider the semantics of plural predicates like 'to be two'. One-place plural predicates can be true of some things: the predicate 'to be two humans' is true of any two humans—for example, Bill and Hillary; 'to be two', any two different things—for example, Washington and Dolly; 'to coauthor a book', any writers that write a book together—for example, Russell and Whitehead; 'to be one', any one thing—for example, Cicero. Now, the plural predication 'Bill and Hillary are two' is true if its predicate 'to be two' is true of Bill and Hillary, to whom the term 'Bill and Hillary' refers. We can extend this account to define the satisfaction and truth of sentences in the fragment of English in question, using plural constructions in the metalanguage, but it is not necessary to do so for the present purpose.

*III.2. Regimentation and logic.* To give a proper treatment of logical relations among sentences in the fragment of English that contains plural constructions, it is useful to regiment the fragment. Regimenting facilitates formulation of logical relations among sentences in it by imposing austere syntax. Let me call the regimented fragment of English the *language of plurals*, and the logic that pertains to sentences in the language (and, thus, their English counterparts) the *logic of plurals.* The language extends elementary language, the regimented language developed in tandem with elementary logic, by including (counterparts of) plural constructions; accordingly, the logic of plurals extends elementary logic to do justice to logical relations involving plural constructions.

*III.2.A. The language of plurals.* The primitive expressions of the language of plurals are:

I. Primitive Terms
    *Singular Constants*: for example, 'Bill' ('b'), 'Hillary' ('h')
    *Singular Variables*: for example, '$\alpha$', '$\beta$'
    *Plural Variables*: for example, '$\alpha s$', '$\beta s$'[22]
II. Term Connective: '⊕' ('AND')
III. Predicates
    *Singular Predicates*
        *logical*: '=' ('is-identical-with')
        *nonlogical*: for example, 'is-a-human' ('H'), 'admires' ('A'), 'is-a-member-of' ('$\epsilon$')

---

[22] Plural variables correspond to plural pronouns as used anaphorically in sentences like 'There are some humans, and *they* cooperate' I put '*s*' to make the variables look like nouns of English in plural forms, but the variables are simple symbols.

*Plural Predicates*
*logical*: 'H' ('IS-ONE-OF')
*nonlogical*: for example, 'COOPERATE', 'WRITE', 'SING-TOGETHER'
IV. Sentential Connectives: '~' ('it-is-not-the-case-that'), '∧' ('and')
V. Quantifiers
*The singular existential quantifier* '∃' ('there-is-something...such-that')
*The plural existential quantifier* 'Σ' ('there-are-some-things...such-that')

The language officially contains only the existential quantifiers and two sentential connectives. But the universal quantifiers '∀' ('anything...is-such-that') and 'Π' ('Any-things...are-such-that') and other sentential connectives that I shall use can be introduced via the usual definitions.

It is not necessary to elaborate on the rules of formation except for some clarifications. First, the plural argument places of predicates admit of any *singular*, as well as plural, term. Thus 'Bill IS-ONE-OF Bill' is grammatically correct, though the second argument place of its predicate is plural.[23] Second, the term connective '**@**' accepts any two (occurrences of) terms to form a complex *plural* term. Thus '[Cicero**@**Tully]' and '[Cicero**@**Cicero]' are plural terms. Third, the first argument place of the plural (or singular) quantifier, marked by '...', admits of only a plural (or singular) variable.

In addition to the predicate '=', familiar from elementary logic, the language of plurals has another logical predicate: 'IS-ONE-OF'. This predicate, as a regimentation of the predicate 'to be one of' in English, counts as a logical predicate for the same reason as the predicate '=' does so: some logical relations rely on the specific nature of the predicate (or its English counterpart). For example, 'Cicero *is one of* Cicero and Caesar' is a logical truth; 'Cicero *is one of* Tully and Caesar' and 'Cicero *is one of* the Roman philosophers' are logically equivalent with 'Cicero is Tully or Caesar' and 'Cicero is a Roman philosopher', respectively. The logical relations can be lost if the predicate is replaced with another.

---

[23] Its first argument place is singular. One might object that the English predicate accepts only a plural term in its second argument place. If so, we can define a new predicate 'is-one-of' with the old predicate:

$$s \text{ is-one-of } t =_{df} \forall\alpha[s \text{ is one of } [t \text{ AND } \alpha]], \text{ where } t \text{ is } \textit{any} \text{ term}$$

and introduce the counterpart of the new predicate as a primitive in the regimented language.

Given the predicate, we can define a plural predicate, whose *i*-th argument place is plural, on the basis of a predicate $\phi$ whose *i*-th argument place is singular as follows:

> $\phi^{P(i)}(t_1,\ldots, t_{i-1}, \alpha s, t_{i+1},\ldots, t_n) =_{df} \exists\beta[\mathrm{H}(\beta, \alpha s) \rightarrow \phi(t_1,\ldots, t_{i-1}, \beta, t_{i+1},\ldots, t_n)]$, where $t_1,\ldots, t_{i-1}, t_{i+1},\ldots, t_n$ are terms appropriate for the relevant argument places of $\phi$ and $\beta$ occurs free in none of $\phi$ and $t_1,\ldots, t_{i-1}, t_{i+1},\ldots, t_n$

Call the predicate $\phi^{P(i)}$ (or simply $\phi^P$ if $\phi$ is one-place) the (*i-th*) *plural expansion* of the base predicate $\phi$. Sentences like 'Bill and Hillary are humans' and 'They are humans' can be seen to be formed by the plural expansion 'is-a-human$^P$' (in symbols, '$H^P$') of the singular predicate 'is-a-human'. We can then explain the logical equivalence between, for example, the plural predication 'Bill and Hillary are humans' and the sentential conjunction 'Bill is a human, and Hillary is a human' as resulting from the definition of the plural expansion; the plural expansion is true of any things (for example, Bill and Hillary) each of which the singular base is true of.

Now, plural definite descriptions like 'the humans' (in symbols, '$(\iota w)H^P(w)$') can be seen to result from applying a description operator to a plural expansion. This description, for example, can be defined in contexts in which it occurs primarily as follows:

$$\phi\llbracket(\iota w)H^P(w)\rrbracket =_{df} \Sigma\alpha s[\forall\beta(\mathrm{H}(\beta, \alpha s) \leftrightarrow H(\beta)) \wedge \phi\llbracket\alpha s\rrbracket]^{24}$$

*III.2.B. The logic of plurals.* I shall now present logical truths in the language of plurals in an axiomatic way.[25] To do so, it is convenient to use schematic letters for different kinds of expressions (or sets thereof) in the language: '$s$' for singular terms; '$t$' and '$u$' for any terms; '$v$' and '$w$' for singular and plural variables, respectively; '$\phi$' and '$\psi$' for sentences; and '$\Gamma$' for sets of sentences. (They are all used with or without numerical subscripts.) Say that $\phi$ is a *plural universal generalization* of $\psi$ if and only if either $\phi=\psi$ or $\phi=\ulcorner Q_1u_1Q_2u_2$

---

[24] $\ulcorner\phi\llbracket t\rrbracket\urcorner$ is used for sentence $\phi$ in which term $t$ occurs primarily There is another kind of plural definite descriptions—for example, 'the coauthors of PM' or 'the things such that something is a human if, and only if, it is one of them' (which 'the humans' can be taken to abbreviate). Such descriptions, too, can be paraphrased into the language of plurals, though somewhat differently 'The coauthors of PM cooperate' can be paraphrased by 'There are some things that are coauthors of PM, they are the same things as any things that are coauthors of PM, and they cooperate'

[25] Although I shall also talk of other logical relations (for example, logical implication), such talks are reducible to talks of logical truth in the usual way

$\ldots Q_n u_n \psi\urcorner$, where $Q_1, Q_2, \ldots, Q_n$ are the singular or plural universal quantifiers whose first argument places admit of the respective variables $u_1, u_2, \ldots, u_n$.[26] Then the axioms of the system are the plural universal generalizations of instances of the following schemas:[27]

*Group A*

I. Truth-functional tautologies
II. $\phi(s) \rightarrow \exists v \phi(v)$, where $s$ is substitutable for $v$ in $\phi(v)$
III. $\forall v[\phi(v) \rightarrow \psi(v)] \rightarrow [\forall v \phi(v) \rightarrow \forall v \psi(v)]$
IV. $\phi \rightarrow \forall v \phi$, where $v$ does not occur free in $\phi$
V. $s = s_1 \rightarrow [\phi \rightarrow \phi_1]$, where $\phi$ is an atomic sentence and $\phi_1$ is obtained from $\phi$ by replacing $s$ in some of its occurrences by $s_1$

*Group B*

VI. $\phi(t) \rightarrow \Sigma w \phi(w)$, where $t$ is substitutable for $w$ in $\phi(w)$
VII. $\Pi w[\phi(w) \rightarrow \psi(w)] \rightarrow [\Pi w \phi(w) \rightarrow \Pi w \psi(w)]$
VIII. $\phi \rightarrow \Pi w \phi$, where $w$ does not occur free in $\phi$

*Group C*

IX. $\mathrm{H}(s, [t \oplus u]) \leftrightarrow [\mathrm{H}(s, t) \vee \mathrm{H}(s, u)]$
X. $\mathrm{H}(s, s_1) \rightarrow s = s_1$
XI. $\exists v \mathrm{H}(v, w)$

*Group D*

XII. $\forall v[\mathrm{H}(v, t) \leftrightarrow \mathrm{H}(v, u)] \rightarrow [\phi \rightarrow \phi_1]$, where $v$ does not occur free in $t$ or $u$, $\phi$ is an atomic sentence, and $\phi_1$ is obtained from $\phi$ by replacing $t$ in some of its occurrences by $u$
XIII. $\exists v \phi(v) \rightarrow \Sigma w \forall v[\mathrm{H}(v, w) \leftrightarrow \phi(v)]$, where $w$ does not occur free in $\phi(v)$

The schemas of *Group A* are familiar from elementary logic.[28] Those of *Group B* are plural counterparts of schemas II-IV. *Group C* contains schemas that concern mostly the logic of the predicate 'H'. The rest are contained in *Group D*.

We can do with one rule of inference, modus ponens: from $\phi$ and $\ulcorner[\phi \rightarrow \psi]\urcorner$, we may infer $\psi$. There are logical truths of the language we cannot obtain by applying the rule repeatedly to the axioms. How can we then distinguish such logical truths from its sentences that are not logically true? I do not give a full answer to this question

---

[26] Recall that the universal quantifiers are defined as follows:

$$\forall v \phi(v) =_{df} \sim\exists v \sim\phi(v); \; \Pi w \phi(w) =_{df} \sim\Sigma w \sim\phi(w)$$

[27] The definitions of '$t$ is substitutable for $u$ in $\phi(u)$' and '$t$ occurs free in $\phi$' (where $t$ and $u$ are variables) are straightforward.

[28] They are taken from Herbert Enderton, *A Mathematical Introduction to Logic* (New York: Academic, 1972), pp. 104ff. (The schema '$s = s$' is derivable given V and X-XI.)

here, because the logic of the language cannot be given a complete axiomatization.[29] For present purposes, it suffices to give a partial answer: the logic is a *conservative* extension of elementary logic, that is:

> Conservation principle: a sentence $\phi$ that belongs to an elementary language (or the singular fragment of the language of plurals) is logically true if *and only if* it is logically true on elementary logic.

Thus, sentences like '$b \neq h \rightarrow \exists\alpha[b\epsilon\alpha \wedge h\epsilon\alpha]$' or '$[\forall\gamma(A(e, \gamma) \rightarrow \gamma = t) \wedge \forall\gamma(A(t, \gamma) \rightarrow \gamma = e)] \rightarrow \exists\alpha[e\epsilon\alpha \wedge t\epsilon\alpha]$' are not logically true even in the language of plurals; they are not so on elementary logic.

*III.2.C. Remarks on the logic.* The conservation principle helps to argue against the prevalent view that to assert plural quantifications like the following, known as the Geach-Kaplan sentence, is to commit one to the existence of sets (or other complex objects):[30]

> (14) *There are some* critics who admire only one another (in symbols, $\exists\alpha s\forall\beta[\mathrm{H}(\beta, \alpha s) \rightarrow C(\beta) \wedge \forall\gamma(A(\beta, \gamma) \rightarrow \beta \neq \gamma \wedge \mathrm{H}(\gamma, \alpha s))]$).

The view rests on paraphrasing such sentences into singular quantifications like the following:

> (15) There is something (for example, a set) every member of which is a critic and admires only its other member (in symbols, $\exists\alpha(\exists\beta\beta\epsilon\alpha \wedge \forall\beta[\beta\epsilon\alpha \rightarrow C(\beta) \wedge \forall\gamma(A(\beta, \gamma) \rightarrow \beta \neq \gamma \wedge \gamma\epsilon\alpha)])$).

But this is not a correct paraphrase; (14) and (15) are not logically equivalent. To see this, consider the following:

> (16) Ezra is a critic, Thomas is a critic, Ezra is not identical with Thomas, Ezra admires only Thomas, and Thomas admires only Ezra (in symbols, $[C(e) \wedge C(t)] \wedge e \neq t \wedge [\forall\gamma(A(e, \gamma) \rightarrow \gamma = t) \wedge \forall\gamma(A(t, \gamma) \rightarrow \gamma = e)]$).

---

[29] The language contains sentences like 'Some positive natural numbers are successors of only one another', which amounts to the negation of the full induction principle that helps to yield a categorical characterization of natural numbers. The logic of languages that contain the logical resources of the language presented above and a two-place predicate (for example, 'is a successor of' or 'admires'), which helps to state the simulacrum of a sufficient fragment of the theory of finite sets, cannot be given a full axiomatic characterization. For a full model-theoretic characterization of the logic, see my "The Language and Logic of Plurals," section IV.

[30] For this view, see W. V. Quine, *Methods of Logic* (London: Routledge, 1972, 3rd edition), p. 239, and his *The Roots of Reference* (La Salle, Ill.: Open Court, 1973), p. 111. See also Michael D. Resnik, "Second-order Logic Still Wild," this JOURNAL, LXXXV, 2 (February 1988): 75-87; and Allen P. Hazen, "Against Pluralism," *Australasian Journal of Philosophy*, LXXI, 2 (June 1993): 132-44. For other, less direct criticisms of the view, see George Boolos, "To Be Is to Be a Value of a Variable (or to Be a Value of Some Variable)," this JOURNAL, LXXXI, 8 (August 1984): 430-49, and his "Nominalist Platonism," *Philosophical Review*, XLIV, 3 (July 1985): 327-44; see also Lewis, *Parts of Classes*

(16) does not logically imply (15), by the conservation principle, because it does not do so on elementary logic. But (16) logically implies the following, which logically implies (14), by schema VI:

(17) Ezra and Thomas are critics who admire only each other (in symbols, $\forall\beta[\mathrm{H}(\beta, [e \oplus t]) \rightarrow C(\beta) \wedge \forall\gamma(A(\beta, \gamma) \rightarrow \beta \neq \gamma \wedge \mathrm{H}(\gamma, [e \oplus t])))]$).

We can see some important features of the logic of plurals through schema VI, which we can use to show that (17) logically implies (14). Sentence (14) can be seen to result from replacing the italicized plural terms in the following sentences with the plural quantifier:

*Ezra and Thomas* are critics who admire only each other (sentence (17)).
*Ezra, Thomas, and Bill* are critics who admire only one another.
*Ezra, Thomas, Bill, and Hillary* are critics who admire only one another.

Thus, by the schema, any of these sentences logically implies (14). Other instances of its antecedent include the following (or their paraphrases):

*John and Carol* are children who carry a piano.
*John and Carol* are children who are healthy.
*Russell, Moore, and Whitehead* are famous humans who are three philosophers.
*Bill, Boris, Tony, and Howard* are *politicians who cooperate*.

Thus, we can use the schema to show that they logically imply the following, respectively:

*There are some* children who carry a piano.
*There are some* children who are healthy.
*There are some* famous humans who are three philosophers.
*There are some* politicians who cooperate.

Notice that sentences like the following, too, are instances of the schema, because the schematic letter '*t*' can be replaced with a singular term as well:

If Cicero is a Roman who is a philosopher, some Romans are philosophers (in symbols, $[R^P(c) \wedge P^P(c)] \rightarrow \Sigma\alpha s[R^P(\alpha s) \wedge P^P(\alpha s)]$).[31]

---

[31] Thus, '$\Sigma\alpha s[R^P(\alpha s) \wedge P^P(\alpha s)]$', '$\exists\alpha[R^P(\alpha) \wedge P^P(\alpha)]$', and '$\exists\alpha[R(\alpha) \wedge P(\alpha)]$', which can be taken to paraphrase 'Some Romans are philosophers', are logically equivalent.

Some might object that these are not logically true, because 'Some Romans are philosophers' implies the existence of at least two Romans. But this is not a correct view. The plural quantification cannot logically imply 'There are at least two Roman philosophers'; it is logically implied by, for example, 'Cicero and Tully are Romans who are philosophers', which is logically implied by 'Cicero is a Roman who is a philosopher, and Tully is a Roman who is a philosopher'.

Schemas XI-XIII help to distinguish the logic of plurals from second-order logic. Schema XII is the plural counterpart of schema V, which states the principle of substitutivity of identity (for singular terms). Schema XII states the principle that, if some things are the same things as some things and the former are so-and-so, then the latter, too, must be so-and-so; its antecedent gives the condition for some things (for example, Russell and Whitehead) to be the same things as some things (for example, the authors of PM). Thus, we can use the schema to show that sentences like the following are logically true:

> If Russell and Whitehead are the authors of PM, and Russell and Whitehead cooperate, then the authors of PM cooperate.
>
> If the children who carry a piano are John and Carol, they carry it if and only if John and Carol carry it.

Consider schema XIII. The following sentences, for example, are instances of its antecedent and consequent, respectively:

(18) There is a philosopher who writes something (in symbols, $\exists\alpha[P(\alpha) \wedge \exists\beta W(\alpha, \beta)]$).

(19) There are some things such that something is one of them if and only if it is a philosopher who writes something (in symbols, $\Sigma\gamma s\forall\alpha(\mathrm{H}(\alpha, \gamma s) \leftrightarrow [P(\alpha) \wedge \exists\beta W(\alpha, \beta)])$).

Thus, by the schema, (18) logically implies (19). Using the schema, we can also show that 'There is a politician in Washington', for example, logically implies 'There are some politicians in Washington such that any politician in Washington is one of them'; this sentence is logically equivalent with the consequent of its appropriate instance.

Notice that sentence (19) can be abbreviated as follows:

(19a) There are *the* philosophers each of whom writes something.

for if some things are such that something is one of them if and only if it is a philosopher who writes something, they are *the* philosophers each of whom writes something (and vice versa). In general, the consequent of schema XIII can be abbreviated as '$\Sigma w_1[w_1 \equiv (\iota w)\phi^P(w)]$',[32] where the predicate '$\equiv$' is defined as follows:

$$t \equiv u =_{df} \forall v[\mathrm{H}(v, t) \leftrightarrow \mathrm{H}(v, u)], \text{ where } v \text{ does not occur free in } t \text{ or } u$$

Thus, we can take the schema to state a principle giving a logically sufficient condition for there to be some things (for example, the philosophers each of whom writes something): if there is something that is so-and-so, there are *the* things each one of whom is so-and-so. Notice that the necessary condition is also sufficient; the converse of the schema holds as well, which we can show using schema XI:

XIV. $\Sigma w \forall v[\mathrm{H}(v, w) \leftrightarrow \phi(v)] \rightarrow \exists v\phi(v)$, where $w$ does not occur free in $\phi(v)$

Thus, the antecedent of XIII is essential to the schema; we cannot do without the antecedent to get a stronger schema of logic. Most instances of its consequent are not logical truths; sentence (19) is not a logical truth. Moreover, the instances include logical falsities, such as '$\Sigma\alpha s\forall\beta[\mathrm{H}(\beta, \alpha s) \leftrightarrow \beta \neq \beta]$', which logically implies '$\exists\beta\beta \neq \beta$'.

One might attempt to preserve schema XIII by treating plural terms and quantifiers as predicates and second-order quantifiers, respectively, because instances of the following schema, the second-order counterpart of XIII, are logically true:

XIIIa. $\exists vX(v) \rightarrow \exists X\forall v[X(v) \leftrightarrow \phi(v)]$, where $X$ does not occur free in $\phi$

But this is not sufficient to reduce the logic of plurals to second-order logic. The converse of XIIIa, that is, the second-order counterpart of XIV, does not hold: '$\exists X\forall\alpha[X(\alpha) \leftrightarrow \alpha \neq \alpha]$', which is logically implied by '$\forall\alpha[\alpha \neq \alpha \leftrightarrow \alpha \neq \alpha]$' on second-order logic, is true; but '$\exists\alpha\alpha \neq \alpha$' is false. We cannot derive the converse of XIIIa, notice, because the second-order counterpart of schema XI (namely, '$\exists vX(v)$') does not hold; '$\exists\alpha\alpha \neq \alpha$' is logically false. Moreover, notice that second-order logic contains a schema stronger than XIIIa:

XIIIb. $\exists X\forall v[X(v) \leftrightarrow \phi(v)]$, where $X$ does not occur free in $\phi$

which we can use to get the negation of '$\forall X\exists\alpha X(\alpha)$', a second-order universal generalization of an instance of the schema '$\exists vX(v)$'. Thus, one cannot take plural quantifiers as second-order quantifiers in an alternative notation.[33]

---

[32] To see this, apply the definition of plural definite descriptions.

It is usual to circumvent this well-recognized discrepancy between the logic of plural quantifiers and the logic of second-order quantifiers by paraphrasing plural quantifiers like 'Some things are such that...' into second-order notation using restricted quantifier phrases like '$\exists X[\exists \alpha X(\alpha) \wedge ...]$'. This scheme of paraphrase is objectionable on syntactic grounds, because applying the scheme to sentences like 'John and Carol carry a piano' or 'A book is written by Russell and Whitehead' fail to preserve the fact that they share predicates with sentences like 'John carries a piano' or 'A book is written by Cicero'. Still, some might defend the scheme by paraphrasing 'A book is written by Cicero', for example, by the second-order sentence '$\exists \beta (B(\beta) \wedge \exists X[\exists \alpha X(\alpha) \wedge \forall \alpha (X(\alpha) \leftrightarrow \alpha = c) \wedge \mathit{write}(X, \beta)])$'. They might argue that this suffices to account for the logic of plurals, if not their syntax, because, for example, this paraphrase, like their paraphrase of 'A book is written by Russell and Whitehead', logically implies their paraphrase of 'A book is written by some things'.

But the second-order scheme cannot preserve the logical relations among plurals encapsulated in schema XII. Consider arguments like the following, whose logical validity is captured by the schema:

(20) Russell and Whitehead cooperate.

(21) Russell and Whitehead are the authors of PM.

Therefore,

(22) The authors of PM cooperate.

To account for their logical validity by applying the scheme to sentences like (20) and (22), it is necessary to appeal to the second-order counterpart of schema XII:

> XIIa. $\forall v[X(v) \leftrightarrow Y(v)] \rightarrow [\phi \rightarrow \phi_1]$, where $X$ and $Y$ are second-order variables, $\phi$ an atomic sentence, and $\phi_1$ obtained from $\phi$ by replacing $X$ in some of its occurrences by $Y$

---

[33] Neither can second-order languages, governed by second-order logic, be taken to provide alternative notations for plural constructions—pace Boolos, who proposes to interpret the formulas in second-order "languages" by rendering them into the language of plurals. He reads the second-order quantifier '$\exists X$' as 'there are some things such that...them...', and argues that this leads to a defense of the logical status of XIIIb—see his "Reading the *Begriffsschrift*," *Mind*, xciv, 375 (July 1985): 331-44, here p. 340. But the plural reading falls short of justifying XIIIb, whose instances, on that reading, include logical falsities.

But this schema cannot be considered a schema of logical truths in languages that contain second-order predicates like '*cooperate*', which one needs to obtain the suggested paraphrases of sentences like (20).[34] Let the second-order predicate '$\mathcal{I}$' indicate the relation of "identity" between properties, and the first-order predicates '$A$' and '$B$' the property of being identical with Russell or Whitehead and that of being an author of PM, respectively. Then '$\mathcal{I}(A, B)$' is false, though both '$\mathcal{I}(A, A)$' and '$\forall\alpha[A(\alpha) \leftrightarrow B(\alpha)]$' are true. Other second-order predicates give rise to the same problem: because the extensional equivalence between, for example, the properties indicated by the predicates '$A$' and '$B$' does not logically imply their "identity," the predications that attribute, for example, the relation indicated by '*cooperate*' to the properties are not logically equivalent even granting their extensional equivalence.

To defend schema XIIa, one might argue that second-order predicates generate opaque contexts. Schema XIIa cannot be taken to warrant substituting 'is Tully' (*salva veritate*) for the co-extensional predicate 'is an author of *Academica*' in the following sentence:

(23) Bill thinks that Cicero is an author of *Academica*.

just as schema VI does not warrant substituting 'Tully' for 'Cicero' in the sentence. But it is hard to appeal to restrictions like this to defend schema XIIa. Sentence (23) cannot be analyzed as a predication composed of 'Cicero' and a first-order predicate (or 'is an author of *Academica*' and a second-order predicate) because 'Cicero'

---

[34] In *Plurals and Events* (Cambridge: MIT, 1993), Barry Schein presents a second-order scheme that does without such predicates. The scheme renders, for example, 'they cooperate' to a singular quantification over events '$\exists e[e$ is a cooperation event $\wedge \forall\alpha(A(e, \alpha) \leftrightarrow T(\alpha))]$', where '$T$' is a second-order variable corresponding to the pronoun 'they' and '$A(e, \alpha)$' amounts to, roughly, '$\alpha$ is an agent of $e$'. This scheme is subject to a version of the so-called Russell's paradox. The scheme renders '$X$s are as many as $Y$s' to '$\exists e[AMA(e) \wedge \forall\alpha(A(e, \alpha) \leftrightarrow X(\alpha)) \wedge \forall\alpha(O(e, \alpha) \leftrightarrow Y(\alpha))]$', where '$AMA(e)$' and '$O(e, \alpha)$' amount to, roughly, '$e$ is an event whose agents are as many as its objects' and '$\alpha$ is an object of $e$', respectively. Thus, Schein must take the following, which amounts to the true sentence 'Any things are as many as themselves', to hold:

$$\forall X(\exists\alpha X(\alpha) \rightarrow \exists e[AMA(e) \wedge \forall\alpha(A(e, \alpha) \leftrightarrow X(\alpha)) \wedge \forall\alpha(O(e, \alpha) \leftrightarrow X(\alpha))]$$

This leads to a contradiction, assuming '$\exists\alpha{\sim}AMA(\alpha)$'. Its instance obtained by the predicate '$F$' defined as follows leads to a contradiction, because '$\exists\alpha F(\alpha)$' holds:

$$F(\beta) =_{df} {\sim}\exists X[X(\beta) \wedge AMA(\beta) \wedge \forall\alpha(A(\beta, \alpha) \leftrightarrow X(\alpha)) \wedge \forall\alpha(O(\beta, \alpha) \leftrightarrow X(\alpha))]$$

'$[AMA(e) \wedge \forall\alpha(A(e, \alpha) \leftrightarrow F(\alpha)) \wedge \forall\alpha(O(e, \alpha) \leftrightarrow F(\alpha))]$', which implies '$\forall X[\forall\alpha(A(e, \alpha) \leftrightarrow X(\alpha)) \rightarrow \forall\alpha(X(\alpha) \leftrightarrow F(\alpha))]$', implies '$[F(e) \leftrightarrow {\sim}F(e)]$'

(or ‘is an author of *Academica*’) occurs in an opaque context, generated by ‘thinks’, in (23); but the second-order sentences considered in the above paragraph are all second-order predications. Some people might perhaps argue that opaque contexts are more prevalent, that they are generated by all second-order predicates (except first-order quantifiers),[35] but this defense renders the schema useless in accounting for the logical relation in question among sentences like (20)-(22).

### IV. VARIETIES OF PLURAL RELATIONS

I shall now discuss various kinds of plural relations, with concentration on plural properties. In particular, I shall analyze a special kind of plural properties, the pure numerical properties, in terms of logical resources only, and confirm the intuitions to which I have appealed.

Being two, recall, is instantiated by some things, but not by anything. It is the same with most of the plural properties discussed above. It is useful to focus on such properties in arguing against the principle of singularity. But, of course, there are other kinds of plural properties. Being one (or being one human), for example, is instantiated by something, for example, Bill, but not by any two or more things (as such). And there are properties instantiated by some one thing as well as by some two or more things (as such): for example, not being two, or *writing a book*, a property instantiated by Cicero and also by Russell and Whitehead (as such). Finally, some plural properties are not instantiated at all; as the singular property of being a Martian philosopher is not instantiated by anything, the plural property of being two Martian philosophers is not instantiated by any one or more things.

Next, recall the plural expansions like ‘is-a-human$^P$’ (in symbols, ‘$H^P$’) of singular one-place predicates like ‘is-a-human’. The plural expansions indicate plural properties that I call the ***plural expansions*** of their ***singular bases***, the properties indicated by the singular predicates. For example, the property that I call ***being human[s]***, indicated by ‘is-a-human$^P$’, is the plural expansion of being a human. Now, any plural expansion is *distributive*, that is, it is instantiated by some things as such if and only if it is instantiated by every one of them.[56] We can see this by considering their analysis in terms of their singular bases. The following definition of the predicate 'being-a-human$^{P}$', for example, can be seen to give an analysis of the property of being human[s]:

---

[35] Some case can be made of this—‘the author of *Academica*’ cannot be substituted for ‘Cicero’ in ‘$\mathcal{I}$ (is Cicero, is Cicero)’ (Notice, however, that the language of plurals officially contains no definite descriptions.) But the problem with Schema XIIa arises primarily because it differs from the correct formulation of the principle of substitutivity for predicates:

$\mathcal{I}(X, Y) \rightarrow [\phi \rightarrow \phi']$, where $\phi'$ is obtained from $\phi$ by replacing $X$ in some places by $Y$

$$H^{P}(\alpha s) =_{df} \forall\beta[\mathrm{H}(\beta, \alpha s) \rightarrow H(\beta)]$$

That is, for some things to be human[s] is for them to be such that every one of them is a human. Thus, the plural property is instantiated by, for example, Bill and Hillary (as such) if and only if it is instantiated by Bill and by Hillary; for '$H^{P}([b \oplus h])$' and '$H^{P}(b) \wedge H^{P}(h)$' are logically equivalent. But this logical equivalence relies on the nature of the plural expansion; sentences of the forms '$\phi(t \oplus u)$' and '$\phi(t) \wedge \phi(u)$' are not in general logically equivalent. Accordingly, most of the other plural properties, such as being two, are nondistributive.[57]

Let me now focus on a special kind of plural properties: numerical properties. I distinguish two kinds of them: *impure* ones like being two humans or being two-membered; and *pure* ones like being two. I shall analyze the impure ones in terms of the pure ones, and the pure ones in terms of logical resources only.

Being two humans, for example, can be analyzed in terms of (i) being two and (ii) being humans[s] according to the following definition:

$$\text{BE-TWO-HUMANS}(\alpha s) =_{df} \text{BE-TWO}(\alpha s) \wedge H^{P}(\alpha s)$$

Thus, for some things to be two humans is for them to be two and also humans. The analysis separates the numerical component of the impure numerical property from its non-numerical component. This clarifies that Bill and Hillary's being of the same kind is irrelevant to their being two.

Similarly, we can distil the numerical component of being two-membered, for example, by analyzing it as follows:

$$\text{is-two-membered}(\alpha) =_{df} \Sigma\beta s(\text{BE-TWO}(\beta s) \wedge \forall\delta[\mathrm{H}(\delta, \beta s) \leftrightarrow \delta\epsilon\alpha])$$

For something to be two-membered is, roughly, for it to have as its members some things that are two. This analysis conforms to, and clarifies, the idea that, for example, the set {Bill, Hillary} instantiates being two-membered *because* Bill and Hillary as such instantiate being two, not vice versa. The relation of membership, indicated by the predicate '$\epsilon$', plays only the subsidiary role of relating the set to the people in question, its members. Thus, the relation, too, is extrinsic to the numerical feature proper of being two-membered.

---

[56] I use 'some things', and so on, to mean the same as 'some one or more things', and so on. Thus, the above sentence implies 'it is instantiated by something if and only if it is instantiated by every one of it'.

[57] There are exceptions like *being at least one*.

It is the same with other impure numerical properties like having one offspring, having two parts, or having three elements. They inherit their numerical character from the pure numerical properties that underlie them as their common components.

Next, we can analyze the properties of being one and being two as follows:

$$\text{BE-ONE}(\alpha s) =_{df} \exists\beta\forall\delta[\mathrm{H}(\delta, \alpha s) \leftrightarrow \delta=\beta]$$
$$\text{BE-TWO}(\alpha s) =_{df} \exists\beta\exists\gamma(\beta\neq\gamma \wedge \forall\delta[\mathrm{H}(\delta, \alpha s) \leftrightarrow \delta=\beta \vee \delta=\gamma])$$

It is straightforward to give similar analyses of the other properties corresponding to natural numbers (greater than one).[38]

Now, we can confirm the logical equivalence of facts like (1) and (3), which can be stated by the following sentences:

(1a) BE-TWO $([b \oplus h])$

(3a) $b\neq h \wedge [\exists\beta\beta=b \wedge \exists\beta\beta=h]$

These sentences are logically equivalent, given the definition of the predicate 'BE-TWO', because the following is a logical truth:

$$\forall\delta[\mathrm{H}(\delta, [b \oplus h]) \leftrightarrow \delta=b \vee \delta=h]^{39}$$

We can also confirm that the existence of, for example, the set {Bill, Hillary} is not a logical consequence of Bill and Hillary's being two. For sentence (1a) does not logically imply the singular construction '$\exists\alpha[b\epsilon\alpha \wedge h\epsilon\alpha]$', because, as we can see using the conservation principle, sentence (3a) does not do so.

Finally, notice that the definition of, for example, the predicate 'BE-TWO' can be modified as follows, using only bounded quantifiers:

$$\text{BE-TWO}(\alpha s) =_{df} \exists\beta(\mathrm{H}(\beta, \alpha s) \wedge \exists\gamma[\mathrm{H}(\gamma, \alpha s) \wedge \beta\neq\gamma \wedge \forall\delta(\mathrm{H}(\delta, \alpha s) \rightarrow [\mathrm{H}(\delta, \alpha s) \leftrightarrow \delta=\beta \vee \delta=\gamma])])$$

---

[38] Or we can analyze them recursively in terms of their respective predecessors:

$$\text{BE-}(N+1)(\alpha s) =_{df} \exists\beta\Sigma\gamma s(\text{BE-}N(\gamma s) \wedge H(\beta, \alpha s) \wedge \forall\delta[H(\delta, \gamma s) \leftrightarrow H(\delta, \alpha s) \wedge \delta\neq\beta])$$

[39] We can show this using schemas IX-X.

This helps to confirm and clarify the idea that instantiating a pure numerical property is a matter intrinsic to the things in question (for example, Bill and Hillary); whether or not there is any other object (for example, the set {Bill, Hillary}) is irrelevant to their being two.[40]

## V. EPILOGUE

The standard conception of relation is the natural ally of the Fregean conception of language and logic. The Fregean language extends the Aristotelian language by including, among others, predicates with more than one argument place; and the Fregean logic extends the Aristotelian logic to the extended language. The development of the Fregean conception, which helped to make clear the limitations of the Aristotelian conception of language and logic, led to the establishment of the standard conception, which improves on the Aristotelian conception of relation that does not allow the existence of relations even in the usual sense. Similarly, we can regiment plural constructions and formulate their logic as in section III, which helps to clarify the limitations of the Fregean conception and replace the standard conception with a broader conception that allows plural relations. Such work, however, does no more than help to remove blindfolds, without which we can directly see the limitations of the standard conception. It is, after all, two things that are two.

BYEONG-UK YI

University of Queensland

---

[40] On my view, natural numbers are the above-mentioned numerical properties. See my *Understanding the Many* (Ph. D. Dissertation, University of California/Los Angeles, 1995), chapter 4; and "Numbers and Relations."

# COMMENTS AND CRITICISM

## WINGS, SPOONS, PILLS, AND QUILLS: A PLURALIST THEORY OF FUNCTION

Beth Preston's "Why Is a Wing Like a Spoon; A Pluralist Theory of Function"[1] contains misunderstandings of my theory of proper functions and of its relation to the notion she calls *system functions*. I should like to clear some of this misunderstanding away, and at the same time show why system function is not a determinate notion and certainly cannot do the job Preston wants it to do in defining artifact functions.

After citing my *Language, Thought, and Other Biological Categories*[2] (and only that), Preston sums up "function theorists" generally, remarking, "it is altogether common to assume that a theory of artifact function will simply fall out of the general theory of function now being articulated, even though this purportedly general theory of function is based on an examination of biological function" (216). This comes immediately after remarking on my treatment in LTOBC that "[t]he one difference she notes [between tools and other things with functions] is that the proper functions of biological devices, language devices, and representational mental states are not the result of purposive design, as are the proper functions of tools" (215). Now, I thought that *that* particular difference between artifacts and, for example, body organs, was apparently a huge difference, indeed, a seemingly insurmountable difference. Moreover, I believe a unique feature of my discussion of proper function in LTOBC was the care with which I tried to articulate the notions of derived and adapted functions (chapter 2) in an attempt to show exactly how, despite first appearances, human purposes themselves have proper functions, and how a result of this is, first, that all human actions have proper functions and, then, that artifacts have proper functions. The attempt to show exactly how such things as artifacts could be subsumed under the same notion of function as that for language devices and biological devices was one of the central and most difficult tasks of LTOBC. Indeed, Preston engages in an extended critique of my way of doing this, clearly belying her claim that "[t]he nature of artifacts generally, and the nature of their functionality in particular, is taken to be so transparently obvious...that virtually no-

---

[1] This JOURNAL, XCV, 5 (May 1998): 215-54.

[2] Cambridge: MIT, 1984; hereafter LTOBC.

0022-362X/99/9604/191–206

body has bothered to examine it at any length" (215). Daniel Dennett, of course, is another who has examined the nature of artifacts at some length.[3]

Nor was my "purportedly general theory of function...based on an examination of biological function." Let me quote from the introduction and chapter one of LTOBC:

> Proper function is intended as a technical term. It is of interest because it can be used to unravel certain problems, not because it does or doesn't accord with common notions such as "purpose" or the ordinary notion "function." My program is far removed from conceptual analysis; I need a term that will do a certain job and so I must fashion one (*op. cit.*, p. 2).

> I need [the definition of "proper function"] in order to talk about analogies and disanalogies among things belonging to quite diverse categories—body organs, tools, purposive behaviors, language elements, inner representations, animal's signals, customs, etc....the purpose being to make as explicit as possible analogies among categories of things, which analogies had struck me as useful to reflect on...the spirit in which I offer [the definition] to the reader is as a handle by which to grab hold of the analogies.... If the analogies are really there, it will undoubtedly take time and an interchange of ideas before they can be explicated entirely adequately (*op. cit.*, p. 38).

From these passages, and also from the definition I gave, and from my examples which were mainly of language functions in the early chapters of LTOBC, I thought it would be clear that the term 'biological' in my title was used not literally, but broadly or metaphorically. It should be noted as well that the compound term 'proper function' was my own coinage. I intended (as suggested at LTOBC, page 2) *Webster's Dictionary*'s first meaning of 'proper', which coincides with that of the Latin *propium*, meaning *one's own*.[4] Certainly, nothing "judgmental" was meant here (224), and if proper functions are capacities a thing "ought to have" (224), then remember that the sky can also look "like it ought to rain tomorrow." Normative terms are not always evaluative, but can indicate any kind of measure from which actual departures are possible. For example, a numerical average is also a "norm." (These last two points are trifling in the context of Preston's essay. I mention them only because other commentators have misunderstood me in the same ways, resulting sometimes in serious confusions. Clearly, I should have been more explicit before.)

---

[3] See, for example, "The Interpretation of Texts, People, and Other Artifacts," *Philosophy and Phenomenological Research*, supplement to Volume v (Fall 1990): 177-94. See also his "Preston on Exaptation: Herons, Apples, and Eggs," this JOURNAL, xcv, 11 (November 1998): 576-80.

[4] The root is found, for example, in 'property' and in the verb 'to appropriate'. 'Functioning properly' is, of course, a perfectly ordinary English phrase, but there 'proper' tends toward an evaluative meaning.

Preston's first serious error results from the assimilation, under her term 'system functions', of the sort of functions Robert Cummins had in mind in his classic paper "Functional Analysis"[3] to Stephen Jay Gould and E. S. Vrba's[4] "exaptations." The categories "exaptation" and "adaptation," as defined by Gould and Vrba, are exclusive. If a thing has been selected for because it has had a certain effect, that effect is an "adaptation" and is ***therefore not*** an "exaptation." Cummins functions, on the other hand, clearly include functions that a system was selected for performing: "Functional analysis in biology is essentially similar. The biologically significant capacities of an entire organism are explained by analyzing the organism into a number of 'systems'—the circulatory *system*, the digestive system, the nervous system, etc.—each of which has its characteristic capacities" (*op. cit.*, p. 760-61, quoted by Preston, 220). Obviously, many of the capacities of these systems were selected for and Cummins is well aware of that. But it is not ***because*** these capacities were selected for that he calls them "functions."

Having conflated Cummins functions with exaptations by calling both "system functions," Preston now claims that "Millikan's pluralist tolerance for system function should translate into a similar tolerance for exaptation, but in fact she is very critical of this notion, arguing that the vast majority of the phenomena taken to be exaptations are actually adaptations, that is, the functions in question are actually *proper* functions rather than *system* functions" (228). But, of course, the claim that something is an adaptation rather than an exaptation has nothing to do with the question whether it has a Cummins function. All adaptations have (or once had) Cummins functions when they are given analyses in the context of appropriate biological systems. Thus, I remain a staunch pluralist. There are (at minimum) two kinds of functions, Cummins functions and proper functions, and lots of functions belong to both these kinds.

Cummins functions are dispositions of parts of some designated containing system, or simpler dispositions of the whole system which, added together, account for some complex capacity one wants to explain. Exaptations are supposed to be Cummins-style functions where (1) the chosen capacity to be explained is how the system contributes to survival and proliferation of a particular species of organism, and (2) the capacity also happens not to be an adaptation. The crucial question for a theory of exaptations is not, as Preston supposes, how often and how soon exaptations become adaptations, but whether "contributes to the survival and proliferation of a species" is a well-defined notion.[7]

---

[3] This JOURNAL, LXXII, 20 (November 20, 1975): 741-65.

[4] "Exaptation: A Missing Term in the Science of Form," *Paleobiology*, VIII (1982): 4-15

In describing the general form that functional analyses take, Cummins himself mentions flow charts, circuit diagrams, and computer programs. The most common uses we make of these items, however, is to describe how a system is *designed*, that is, quite literally, what *someone* designed it to do. Designed systems come with all kinds of specifications (both explicit and implicit) attached. They come with assumptions about what to count as parts of the system versus parts of the outside world or things accidentally dropped into the system; about what to count as appropriate background conditions in which the system is to operate; about what to count as appropriate input to and output from the system; about what to count as "state changes" within the system as opposed to damage, breakdowns, or weardowns; about what to count as the system itself having a certain capacity versus some interference from outside the system's being partly responsible for something it does; and so forth. Unfortunately, members of living species do not come with directions telling what is relevant to proper operation of the system rather than irrelevant or external to it. To be an exaptation, however, a thing must first of all be a part of the system under analysis. How do we determine what is to count as part of the biological system?

Let me first parody. Cats have an interesting capacity to walk which it would be interesting to have analyzed. But anaesthetized cats have exactly the same capacity; indeed, dead cats do, if strung up as puppets. If only the capacity to walk is in question, it is not logic that rules out attached puppet strings as suitable conditions for function and forces exerted on these strings as "input" while ruling in a surrounding atmosphere containing the right amount of oxygen, the right force of gravity and a surface underfoot that affords enough friction. To define a biological system, it obviously is not enough to indicate a chunk of matter and an interesting capacity that it has. Which among the many ways that various members of the species might be maintained are relevant to a biological system? An obvious beginning answer is that the appropriate conditions and input must characterize or have characterized various actual members of the species. Which actual members?

---

[7] See my *White Queen Psychology* (Cambridge: MIT, 1993), chapter two.

Given conditions that obtain in a modern hospital, the sounds that one's heart makes can easily become an operative part of a set of sufficient conditions leading to quick diagnosis and medical attention for a life-threatening heart ailment, producing helpful "inputs" to one's system via needles and pills. Does this make these sounds exaptations? Similarly, if the heart emits a healthy sound, this may function immediately to focus medical attention somewhere else, failing which one's life might be lost. Do the conditions and inputs associated with modern hospitals count as appropriate ones in which to consider the species-maintaining functions of parts of human bodies? Have enough humans survived yet due to kidney machines to admit their presence and interference as allowable conditions for the human biological system to run in? How about penicillin injections or pills? Compare these conditions and inputs, for example, with what loving parents put into human infants. A carefully structured environment and proper input is necessary for the continued normal functioning of many biological systems. On what principled line then do we distinguish between 'within' and 'without' for the species-maintaining Cummins system. Hence in what principled way do we define "exaptation"?

Nor is this sort of difficulty in delimiting the biological system's boundaries tied just to purposive interactions with such systems. Porcupines are occasionally saved from breaking bones when they fall out of trees (something which they do astonishingly often) by the springiness of their quills. Should we include this helpful disposition of their quills as an exaptation? Porcupines are especially fond of pine-tree bark, hence when they fall out of trees, it is often on soft pine needles. Is their love for pine bark an adaptation to prevent harm when falling? Should we also include as exaptations the dispositions toward religion that sometimes result in bibles in breast pockets that stop stray bullets that would otherwise have killed soldiers? How frequently do such events have to occur in the life of a porcupine or a soldier to be exaptations? Why exactly that frequently?

Flow charts, circuit diagrams, and programs typically describe systems that are purposefully maintained by humans in conditions which will not relevantly alter the materials of which they are made and which are purposefully isolated from all but appropriate inputs. Their analyzed capacities correspond to purposes humans have for them and their analyzing capacities are how humans intend them to operate. For these reasons, it is clear how they should be analyzed as having Cummins functions. A life system maintains itself. Moreover it often does so in an environment that changes in important respects both over the individual organism's lifetime and from generation to generation. The question what counts as part of the "species-maintaining" Cummins system, as opposed to helpful but accidental interference from outside, has no determinate answer.

Systems described by flow charts, circuit diagrams, and programs generally fulfill their assigned capacities deterministically, so that, if the system does not break down, the capacity is always realized. Many important capacities of animals, on the other hand, are stochastic capacities. The hunting behaviors of predators typically enable them to capture prey only some of the time. The eye-blink reflex protects the eye from damage only once in a while, usually being triggered unnecessarily. If we set aside whether a capacity has been selected for, and also whether it is required to help fulfil another capacity that has been selected for, there is no principled way to distinguish among such capacities whether to count them within a species-maintaining Cummins system or not. Nor will it help to require that each member of the continuing species use a stochastic capacity some of the time. The basic needs of the organism are often met in alternative ways. What serves a certain function often is not actually necessary to its accomplishment. A domestic cat may never successfully employ either its ability to see prey in the dark or its ability to fish, making a good living nonetheless by daytime mouse hunting, perhaps, and by charming people. For animals that learn or reason out ways to do things, if one path to "species maintenance" is blocked, another will be taken instead. Many humans have survived in part because they have learned to hang warm clothes from their shoulders, but if they did not have shoulders they would attach the clothing some other way instead. The capacities to see prey in the dark, to be a shoulder clothes hanger, and so forth are thus not necessary to species maintenance. Do they count, then, as part of the Cummins style system of species maintenance? Do shoulders have supporting clothes as one of their species-maintaining Cummins functions or not? If they do not, does the pancreas also not have as a species-maintaining function the production of insulin because in its absence people take insulin by injection?

My conclusion is that there can be no well-defined notion here unless without principle, that is, by arbitrarily stipulating some percentage of individual lives in some stipulated population that must have happened to be furthered by a candidate exaptation. And even so, it is not clear how we are to deal with the things which have sometimes been useful but were nonetheless entirely dispensable. And because there is no such thing as *the* Cummins functions associated with maintenance of a species, there is no such thing as *the* exaptations associated with it.[8]

The only way to place nonarbitrary limits on what counts as part of a biological system is to bound it with adaptations, with proper functions. Only Cummins functions that either are adaptations or support adaptations are sensibly treated as belonging to biological systems. That was the conclusion of my 1993 chapter on exaptations with its "expanded" notion of proper function which Preston decries. She herself ignores all problems concerning delimitation of the biological system. She also ignores the restrictions I placed on these "expanded" proper functions, namely, that they must be functions of reproduced parts of the same organism.[9] I now think Preston is right, however, that it would be better not to treat these expanded functions as varieties of proper function for the other reason she gives. The notion of a proper function was introduced as a tool for understanding purpose and failure of purpose. If an item has been selected for producing some effect, production of this effect can be held up as a measure or norm from which departures are possible. But my expanded functions were not selected for, hence they fail to exhibit this most essential property of proper functions.

Most important, Preston has thoroughly misunderstood my notions of "adapted proper function" and "derived proper function." These are not add-ons to my original definition of proper function, and they are not established "without any history of reproduction" (233) or through some kind of equivocation in the notion of selection. Let me try to explain.

---

[8] For more careful reflections on this matter, see my "Bio-functions: Two Paradigms," in Cummins, A. Ariew, and M Perlman, eds., *Functions in Philosophy of Biology and Philosophy of Psychology* (New York: Oxford, forthcoming).

[9] Preston claims, for example, that my expanded definition of "proper function" will give human noses the function of holding up eyeglasses. Her idea is that eyeglasses, having continued to be reproduced because they helped people see better, have making people see better as a proper function, and a normal condition for performance of this function is the presence in the right place of a human nose. But this ignores the restriction I gave on those normal conditions which were to have proper functions, namely, that they were to result from reproduced structures of the same organism. I suppose it might be possible to define a "species-maintaining" Cummins system for eyeglasses which would require humans in the environment for its operation, but this system would not be a human organism, nor would it be this system which was responsible for reproducing noses. She has not shown, then, that on my definition, it is any more a proper function of my nose to hold up my eyeglasses than it is a function of a dog's stomach to provide shelter for roundworms. For further discussion, see my "Bio-functions: Two Paradigms."

Consider the perceptual, cognitive, and behavioral systems of an organism. These mechanisms, I have argued, have proper functions.[10] These functions are performed by altering the relation between the organism and the environment as needed so that the environment will provide appropriate conditions and inputs for the organism. Some of these functions involve changing the environment to fit the organism, some involve changing the organism to fit the environment, and some involve changing merely spatial relations between organism and environment, or involve some combination of these three. The characteristic systems that are employed in coordinating states and happenings between organism and environment have relational proper functions of one kind or another. That is, their job is to make it the case that the organism and the environment bear some particular relation to one another. Their job is to create relational structures.

As with any other proper function, a relational proper function of a mechanism corresponds to an effect which ancestors of the mechanism have historically had which accounted for their selection. In this case, the effect was the creation of an abstract relational structure. Consider a photocopier. Does it do the same thing every time you use it? One time it turns out a manuscript on Gottlob Frege, the next time a picture of Marilyn Monroe. But the important thing is that it turns out, every time, something that looks like your original. In this sense, it does the same thing every time. Its effect is always the production of *the very same abstract relational structure*, namely, sameness between the pattern of surface marks on what is in its scanner and what is in its output tray.

In order to produce this relational structure, to effect that this relation exists, it does not need to produce both relata. It produces only one of the relata, but in such a fashion as to effect the existence of the designated relation. Nor is there any cheating here. No device can produce an effect, a result, a product, *ex nihilo*. Every function performed is performed using materials of some kind, using properties of materials already at hand. Producing a relational structure by being guided by one relatum to produce the other is as legitimate a form of producing as producing a wooden spoon by choosing a piece of wood and carving it.

The example I originally used of a mechanism with a relational proper function was the chameleon's pigment rearranging mechanism (LTOBC, chapter two). Its function is to produce the relational structure *skin-bearing-the-same-color-as-its-background* for the chameleon, a further proper function being, of course, to prevent predators from seeing the chameleon. To create this relational structure, the mechanism effects changes in the chameleon but not in the environment. Other animals effect production of similar relational structures by moving into parts of the environment which match them, that is, by changing the spatial relation between themselves and the environment. And likely there are animals that change the environment, say, the surroundings of their nests, in order to produce this kind of relational structure.

---

[10] *White Queen Psychology*, chapter two.

There are a great number of relational structures besides those involving sameness relations which mechanisms may have as proper functions to produce. I have argued that any mechanism whose job is to produce a representation of fact (or to produce an "indicative intentional icon"—a more general category than that of representations of fact) has as a proper function to produce a relational structure (LTOBC, chapter six). An accessible example that I keep repeating is the mechanisms that produce the dance of the honey bee. The job of these mechanisms is to produce a relational structure having a location of nectar as one relatum and a certain aspect of the pattern of the dance as the other. The relation in question is the one given by the abstract function (mathematical sense of "function") describing the semantic rules for the B-mese used by the particular species of bees. In most species, an angle that the dance pattern movement marks out relative to aspects of the hive always bears the same definite relation to the angle of the location of nectar relative to hive and sun. The properly functioning dance mechanism always produces exactly the same thing, namely, this designated relational structure.

I have also argued that any mechanism whose job is to produce a representation of a goal (or to produce an "imperative intentional icon"—a more general category than that of representations of goals) has as a proper function to produce a relational structure (LTOBC, chapter six). An accessible example, once again, is the bee dance, which is both indicative and imperative.[11] The mechanisms that produce bee dances have as a further proper function, beyond producing the dance, to send watching fellow worker bees off in a certain direction. The relational structure thus produced has as relata (1) the angle that the dance pattern marks out relative to aspects of the hive and (2) the angle in which the watching bees fly relative to the hive and the sun. In this case, the mechanisms create this relational structure by making changes not in the organism itself but in its environment—changes not in the bee itself but in the bee's fellow workers. As a result of producing, first, the proper dance/nectar-location structure, the properly functioning dancing-making mechanism later effects always exactly the same thing, namely, existence of this second relational structure between angle of dance and angle of flying workers. (This requires the right environment, of course, one in which there are well-functioning fellow workers available. Their presence is a historically "normal condition" for managing to perform this function.)

---

[11] It is a "pushmi-pullyu representation." See my "Pushmi-pullyu Representations," in James Tomberlin, ed., *Philosophical Perspectives*, Volume IX (Atascadero CA: Ridgeview Publishing, 1996), pp 185-200. Reprinted in *Mind and Morals*, L. May and M. Friedman, eds. (Cambridge: MIT, 1996), pp. 145-61.

Where the proper function of a trait is to produce a series of effects each causing the next it is also the proper function of each stage, on the definition of proper function that I gave, to produce the next stage. So what we have here is the production of one relational structure—dance/nectar-location—having as a proper function the production of another—dance/direction of flying. Finally, and as a logical result (given Euclidean geometry), one more relational structure is produced, namely, workers flying toward nectar. This is typical. Proper-functional relational structures typically do their jobs by producing further relational structures that eventually produce a relation between organism and environment yielding conditions or inputs that the organismic system needs. Sometimes, the whole process involves changing only the environment, and sometimes, only the organism; but sophisticated relational proper functions typically involve both. They also typically cooperate with other structures having other relational proper functions, such as the bee dance producers' cooperation with answering mechanisms in fellow worker bees, which mechanisms could be given a similar relational analysis.

The general picture, then, is of proper-functional processes that involve series of interweaving stages each of which is an abstract relational structure, some moments in these processes producing changes in the organism, others producing changes in the world, but always involving the exactly the same relations, though among different relata each time they are run. They are reproduced invariant processes, always the same when described in the way that explains how they work, that is, what their relevant Cummins functions are, yet different in their elements each time. It was to simplify the description of these complex relational structures and processes that

I introduced in LTOBC the terminology 'adapted proper function' and 'derived proper function'. I intended this merely as useful nomenclature. It is not an addition or a set of extra clauses widening or narrowing the original definition of 'proper function', but merely a way of talking more easily about phenomena that had already been captured by that notion, given that traits and mechanisms can have relational proper functions.

When a mechanism has a relational proper function, it may produce one of the relevant relata while the other relatum is not affected but either (1) remains the same or (2) undergoes its own independent course of development. A simple example of (1) is the chameleon's pigment-arranging mechanism which changes the chameleon's color while the color of the background remains the same. A simple example of (2) I take from Brian C. Smith.[12] Suppose that a species of sunflower not only tracks the sun, but continues to move when the sun goes behind a tree so as to catch up with it on the other side. Here, the mechanism produces changes in the organism designed to maintain a certain organism-environment relation, and does so as the environmental relatum is changed, but not, of course, by the flower. Similarly, the relational structure that consists in the retinal image of a male hoverfly mapping the position and angle of approach of a passing female has as an eventual proper function production of another relational structure consisting in the male's intercepting the female.[13] As with the sun and the sunflower, the female is (as yet) unaffected by the male. Only a change in the direction of flight of the male is effected.

Now examine the relatum that is actually *produced* by a mechanism with a relational proper function. Consider the brown skin produced by the pigment arrangers of the chameleon sitting on a brown background. The job of the pigment arrangers is to produce the relational structure, *skin-color-matching-its-background*. But neither relatum, brown background or brown skin, is an operative part of an historically normal set of sufficient conditions explaining the capacity of the chameleon to become camouflaged. Either relatum might have been replaced, and so long as the other was similarly replaced, the chameleon would have been properly camouflaged. Being brown is not an operative part of a normal set of sufficient conditions for performance of any of the chameleon's functions. Neither the relatum produced nor the relatum to be matched has a proper function nor, notice, any kind of Cummins function either. Not all by itself! Only the whole relational structure has either kind of function.

---

[12] *The Origin of Objects* (Cambridge: MIT, 1996).

[13] For details, see my *White Queen Psychology*, chapter 11.

*Given that brown is the color of the background,* however, the job of the pigment rearrangers is certainly to make the skin brown. I call this kind of job an *adapted proper function* of the mechanism. Turning the skin brown is not usually a proper function of the mechanism, and it will not remain a proper function of the mechanism when the chameleon no longer sits on something brown. Right now, however, given that it is on a brown background, it is an adapted proper function of the mechanism to make the chameleon brown. Turning the skin brown is a proper function of the mechanism "as adapted to" the brown color underneath. It is not, of course, a simple, but only a conditional proper function of the mechanism, an "iffy" proper function, but the 'if' part has been asserted.

The product produced by a device performing an adapted proper function I called an *adapted device.* The chameleon's brown color is an adapted device. It is not, of course, brown itself that is an adapted device, but only the brown skin color of a chameleon produced in the right way in the right circumstances. Thus, Preston is certainly right that we do not want to move, for example, from the premise that a certain English word token has a function derived from its particular speaker's intention to the conclusion that the English language has changed (237). But the idea that derived proper functions are, as she rather misleadingly puts it, "established for particular exemplars" clearly is not "inconsistent" with the notion that all proper functions "essentially involve a selection history" (234).

Now ask about the functions of the relata themselves in a proper-functional relational structure. The proper function of a whole relational structure, as was said, is often to produce another relational structure. The bee-dance mechanism produces the relational structure, *dance-mapping-the-location-of-nectar,* a proper function of which is eventually to produce another relational structure, *worker-bees-heading-toward-nectar.* The nectar, of course, is and remains an independent relatum (nor, of course, is it reproduced by the biological system), so taken by itself, it cannot have a proper function derived from that mechanism. But the other relatum, the bee dance, is produced by the mechanism in accordance with an adapted proper function: it is an adapted proper function of the mechanism, as adapted to the location of the nectar, to produce a certain bee dance, one that maps this location. Further, it is an adapted proper function of the dance-producing mechanism to produce, as a result, a certain direction of flight in fellow worker bees. *Is it also* a proper function of the dance itself to produce this direction of flight?

The answer may at first seem to be "no," for it seems theoretically possible, at least, that the particular bee dance has no ancestors. Perhaps no bee in the bee's lineage ever danced this particular dance before, because there never happened to be nectar located in this particular direction from their hives before. Then this particular bee dance, having never occurred in the past, certainly could not have been selected for any effects that it had, hence could not possibly have any proper functions.

But this overlooks a principle that is fundamental. Because bee dances that map different directions are different from one another does not mean they are not also the same. The original and the photocopy may be different colors because colored paper was used for one but not the other, but the manuscript about Frege, or the picture of Marilyn, is the same. What is of interest is whether there is *a* sameness among the dances such that they are all able to do *something* that is the same, and whether that very something is what their ancestors were selected for doing.

The various dances of a given species of bees are very much the same—indeed, to an untutored observer, hardly discriminable. And when they function in the way that has accounted for the natural selection of their producers and of their answering mechanisms in other workers, they always do exactly the same thing. They produce a direction of flight that is a given function (mathematical sense) of certain aspects of their form—in every case exactly the same function of that form. In LTOBC, I put this rather awkwardly by saying that they have ("direct") proper functions which are adapted to their own concrete forms (LTOBC, chapter two and elsewhere). I am not sure that I have put it less awkwardly here. But the phenomenon itself is really quite easy to grasp. The bee dance has been selected in part for its capacity to cause other worker bees to be guided in their direction of flight *by its form*. In this respect, all bee dances (of the same species) have exactly the same proper function. Strictly speaking, this function is a relational proper function. The dance's job is to cause the workers to fly in a direction that bears a certain relation to itself (for example, LTOBC, p. 42).

Because of this relational proper function, depending on the particular form of the bee dance, it has as an adapted proper function to cause worker bees to fly in some particular direction—say, south-south-west—just as the relational proper function of making the chameleon's skin match its background results in an adapted proper function to turn the skin brown when the chameleon is sitting on something brown. I have called this kind of function a *derived proper function*, derived originally from the proper function of the dance-producing mechanism. It is derived from the particular form of the bee dance which is itself an adapted device, derived from the adapted function of the dance-producing mechanism as operating in a certain context. All derived proper functions are adapted in the same sort of manner.

In LTOBC, I generalized this idea:

> The proper functions of adapted devices are derived from proper functions of the devices that produce them that lie *beyond* the production of these adapted devices themselves. I call the proper functions of adapted devices "derived proper functions" (LTOBC, p. 41).

I was not careful to distinguish in this particular passage between proper functions of adapted devices that are relational but not derived from the producer's particular adaptor (for example, the function common to all bee dances) and those that are not relational and are derived from the producer's particular adaptor.[14] I had in mind the latter kind. Indeed, an adapted device *described qua adapted* is different from other adapted devices derived from the same producer but in different adapting circumstances. And I emphasized that the interesting thing about derived proper functions is that things that have them can be things that are "quite new under the sun." That is, I meant things with adapted proper functions as being named or described in accordance with their adapted aspects (they point in different directions, they are different colors) not the aspects that make them like their ancestors (they are all bee dances, they are all pigmented skins of chameleons). Things with derived proper functions, then, are individual things, not lineages or types of things. It is the brown color of this chameleon that has a derived proper function, not brown itself, indeed, not even brown chameleon skin color, for maybe someone painted a chameleon brown.

Against this background, I argued in LTOBC that there could be relational proper functions that produced adapted devices themselves having relational proper functions, producing more adapted devices having further relational proper functions, to any degree of nesting, and that out of this sort of structure could come things (tokens) seemingly very new indeed under the sun but which still had derived proper functions. I argued that in this sort of way, human desires have derived proper functions derived from the mechanisms designed to form concepts (adapted to the individual's experience[15]) and then to form desires (adapted to other aspects of the individual's experience), and that the derived proper function of a desire is to get itself fulfilled (a relational function). (Reminder: that a thing has a proper function does not imply that it is likely to perform that function!)

---

[14] I distinguished these in the pages immediately following, however.

Further, if the desire is to produce a certain result by means of making, for example, a certain tool or producing a certain sentence, then a derived proper function of the tool (token) or the sentence (token) is to produce this result. Thus, it happens that artifacts have as derived proper functions the functions intended for them by their makers (contrast with Preston: not their users). I also argued that learned behaviors have the results that effected their having been learned as proper functions, direct proper functions when the learning is by trial and error, otherwise derived.

Notice, by contrast, that Preston's analysis of artifact function has the peculiar result that artifacts have no functions except in the sense that they may function *as*, say, a hammer or a screwdriver in someone's hands, and no function unless they have actually been reproduced on account of serving this function. So, if I carefully design and make a new sort of implement for the purpose of opening cans, but nobody ever actually uses it, there is *no sense whatever* in which it is a can opener, or in which its function is to open cans. But surely its purpose, at least, is to open cans, and purpose is what proper functions were originally supposed to cast light on. My claim in LTOBC was that artifacts can acquire proper functions either in Preston's way, that is, "directly," or through the maker's (not the user's) intentions, or in both ways. I also emphasized the possibility not only of shifts in proper function but also of conflicts in function when items have more than one source of function.

It should be clear from the above discussion that the way one describes the things that have proper functions, and the way one explains how a derived function is derived, makes a great deal of difference. Sloppy descriptions of these matters will get one into a lot of trouble—as they have Preston. She herself claims that in talking about these matters, the descriptions under which one considers various functional items do not matter because "there is a whole range of levels of abstraction on which such explanations may move, and which level is appropriate depends on our specific explanatory interests" (235). This is a serious misunderstanding. What counts as a correct description of a proper-functional trait and what counts as the historically normal explanation for its proper functioning, hence as the historically normal conditions for its proper functioning, are, ontologically, completely determinate matters, regardless of how difficult the epistemological issues become. These descriptions and explanations concern the actual historical selection pressures and historical causal mechanisms that produced the effects selected for in successful ancestors. The description of these causal mechanisms has to be one that applies univocally to all of the ancestors. There cannot be separate descriptions for separate ancestors, or there is no proper function but merely an accidental change in the gene pool. Indeed, it also matters very much how traits and mechanisms are described in order to ascribe Cummins functions to them—something that Preston overlooks as well.[16]

---

[15] For details, see my "A Common Structure for Concepts of Individuals, Stuffs, and Basic Kinds: More Mama, More Milk, and More Mouse," *Behavioral and Brain Sciences*, XXII, 1 (February 1998): 55-65; also in E. Margolis and S. Laurence eds., *Concepts: Core Readings* (Cambridge: MIT, 1999) Also see my "Author's Response: With Enemies Like This I Don't Need Friends," *Behavioral and Brain Sciences*, XXII, 1 (February 1998): 89-100, and my *On Confused Ideas* (New York: Cambridge, forthcoming).

RUTH GARRETT MILLIKAN

University of Connecticut

---

[16] For details, see "Bio-functions: Two Paradigms."

# BOOK REVIEWS

*Realistic Rationalism.* JERROLD J. KATZ. Cambridge: Massachusetts Institute of Technology Press, 1998. xxxiv + 226 p. Cloth $35.00.*

Although billed as a book in the philosophy of mathematics, Jerrold Katz's *Realistic Rationalism* is the final chapter in his larger philosophical project of rebutting contemporary naturalism and regaining for metaphysics a central place in philosophy. Relying upon his earlier *The Metaphysics of Meaning*[1] and *Language and Other Abstract Objects*[2] for some arguments against W. V. Quine's and Ludwig Wittgenstein's cases for naturalism, Katz takes positive steps in this work toward formulating and justifying a new realist ontology and rationalist epistemology. He initiates this task as an investigation in the philosophy of mathematics, where the tension between realism and naturalism comes into sharp focus, but he soon broadens his discussion to defend the existence of synthetic a priori knowledge in mathematics, science, and metaphysics and to limit the scope of skepticism. Its wide range and novel ideas make this a rich and stimulating book.

Katz defines realism as the view that abstract entities exist, and sees it as facing three major objections: (1) by their very nature abstract entities are unknowable; (2) determinate reference to them is impossible; (3) the abstract and concrete distinction is untenable. He answers the first two objections by arguing vigorously against the naturalist presuppositions to which its recent advocates have appealed, and by showing how rationalism can provide realism with a respectable companion epistemology. In what I found to be the most interesting part of the book, he responds to the third objection by introducing a new category of objects that bridge the gap between the abstract and concrete. These composite objects—the equator and linguistic tokens are examples—contain both abstract and concrete components. Unfortunately, I shall have the space here only to discuss Katz's responses to (1) and (2).

Katz bases part of his argument for realism and rationalism upon the inadequacies he finds in nominalism and empiricism. For example, he claims that Hartry Field's nominalism breaks down when

---

* I am grateful to Jerrold Katz and Benjamin Theis for comments on an earlier draft of this review.

[1] Cambridge: MIT, 1990.

[2] Totowa, NJ: Rowman and Littlefield, 1981.

0022-362X/99/9604/207–11

one tries to extend it to linguistics, since it confronts the "problems with inscriptional nominalism, which doomed the Goodman and Quine enterprise" (12). On the other hand, Quine's empiricism fails, he argues, because it cannot account for the certainty of mathematics and logic. The upshot of this criticism is that we must recognize that mathematics, linguistics, and logic (collectively the formal sciences) require an ontology of abstract objects and this, in turn, requires a rationalist epistemology. This epistemology should be based on reason alone. It must shun both Platonic perceptual acquaintance with abstract objects and Kantian sensuous intuition so that it is entirely independent of experience, whether sensory or introspective.

An epistemology of this type will skirt Paul Benacerraf's famous point that it is hard to see how a satisfactory epistemology could accompany a realist account of mathematical truth, since this presupposes that the only satisfactory epistemologies are empiricist. Moreover, we are not bound to such an epistemology for the formal sciences, according to Katz, for no one has established empiricism as the best overall theory of knowledge, and it cannot account for certainty in the formal sciences in any case.

Reason need not generate our beliefs about abstract entities, but, in Katz's view, it alone must be used to justify them. For what counts as adequate grounds for a proposition depends upon its subject matter; and if it is about abstract entities, then its justification must be nonempirical. Abstract objects have their intrinsic properties and relations necessarily, and we can never know these through experience, since it can never tell us what must be. (This gives realism the resources to explain the necessity of truths in the formal sciences. They are necessary because they are about abstract objects and describe the unchangeable properties and relations of unchangeable objects.)

How then do we justify knowledge about abstract entities? Katz answers that we do it through proof, rational intuition, and systematization. A proof grounds a proposition about abstract objects by demonstrating the impossibility of the objects' being other than as the proposition claims. But, of course, this works only if the premises of the proof are based upon reason alone. Some of these, such as the basic "ratiocinative knowledge" that four is composite, are based upon rational intuition, while others, such as the principle of mathematical induction, result from systematization. The latter is a process through which we build up systematic connections between clear and correct insights and intuitions, and use the resulting body of theory to correct unclear intuitions.

I shall return to Katz's epistemology for abstract entities, but let us now pass on to reasons for thinking that reference to them is determinate. According to Katz, all the arguments to the contrary begin by excluding certain methods for fixing the reference of a word and then describe a situation in which the admissible factors determining the reference of certain terms are symmetric with respect to a certain class of objects. Thus, if the only constraint on whether the members of certain sequence are the references of the numerals '0', '1', '2', and so on is that the sequence be a model of the Peano axioms, then, of course, it will remain indeterminate whether the numerals refer to the Zermelo numbers rather than the von Neumann numbers. But, according to Katz, we can break this symmetry: the numerals do not refer to sets, for numbers are not sets; "sets but not numbers have members, have subsets, contain the null set, and so on, while numbers but not sets can be prime or composite, even or odd, perfect or imperfect, and so on" (104). Katz does not tell us how we know this but I presume that we know it through rational intuitions. For we cannot prove that numbers are not sets, and I do not see how we could arrive at this result through systematization without begging questions.

Benacerraf's famous argument, which Katz is countering here, assumes that the mathematician's interest in the numbers stops at the level of structure and that it is a mistake for philosophers to try to go deeper. But Katz replies that Benacerraf not only has given no argument for these assumptions, but also falsely assumes that the essential properties of numbers are structural. In fact, some of their properties—for example, their abstractness—are nonstructural. Thus, Benacerraf has illegitimately excluded common-sense and philosophical considerations that would fix the reference of number words. This lapse is due, according to Katz, to Benacerraf's acceptance of the Quine-Wittgenstein view that philosophy has no subject matter of its own and the Fregean view of sense as the determiner of reference.

If we shed these doctrines, and also accept Katz's theory of meaning (developed in *The Metaphysics of Meaning*), then we will admit an important role for synthetic a priori knowledge. For since our conception of the analytic will be much narrower than Gottlob Frege's and more like John Locke's, much mathematics, science, and philosophy will count as synthetic and a priori. Proof, rational intuition, and its systematization will provide the ingredients for an epistemology for these truths.

Clearly, the most controversial part of this epistemology is its appeal to intuition, so I am going to let Katz speak for himself:

> [An intuition is] an immediate, i.e., noninferential purely rational apprehension of the structure of an abstract object, that is, an apprehension that involves absolutely no connection to anything concrete.... What is present to our minds in a clear and distinct intuition of abstract objects is the fact that their structure puts the suppositions of their being otherwise than as we grasp them to be beyond the limits of possibility.... The rationale for claiming that it is intuition that is the source of basic mathematical and other formal knowledge is...There are cases in which we can eliminate everything but intuition as a possible explanation of how it is known that a premise or step in a proof has no counterexample (44-45).

If you find intuition mysterious, as I do, Katz cautions us to distinguish the mysterious from the mystical, and reminds us that philosophy is full of mysteries. Indeed, he is confident that there is "no reason whatsoever to think that scientific psychology cannot in principle explain the rational faculties on which our account of mathematical knowledge is based, though, to be sure, current scientific psychology is far from telling us much about them" (116).

I am less sanguine, but I do not have space here to give my reasons. In any case, before I could discuss this issue with confidence, I would want a fuller and clearer account of Katz's theory of rational intuition than I found in his book. One problem I have traces to his distinction between basic and transcendent knowledge of abstract objects, with the former being obtained through intuition and the latter through systematization. Suppose that we try to apply this idea to the law of trichotomy, which figures as a premise in an example proof Katz presents (40). This is a general law, thus it appears transcendent, and Katz seems to take it to be one whose necessity is known to us. It also seems to be exactly the sort of nonlogical principle that forges connections between pieces of basic knowledge. How is it known? By systematization, as Katz suggests general principles are known? But then how does systematization get off the ground? Maybe we can just hypothesize some initial generalities and plan to revise them if we find counterexamples or they lead to unsatisfactory systems. (This is consistent with Katz's view of the a priori; see section 2.6.3.) On the other hand, perhaps we know tricotomy by intuition, since Katz writes concerning another principle, "even mathematically naive people immediately see that, if m things are put into n pigeon-holes, then when m is greater than n, some hole must contain more than one thing.... The only remaining explanation for the immediate knowledge of the principle is intuition" (45). Let us suppose, then, that intuition can supply us with enough general laws to get systematizing going. Katz characterizes intuition (in the quote above) as an "apprehension of the structure of an abstract object." This would account for our ability to intuit the number four and see that four is composed of two and two. For the pigeon-hole principle we would need more. Katz has written me that in this case our knowledge is based upon "a concatenation of intuitions." I would like to know more about this, and why it will not supply us knowledge of mathematical induction, which, as we have noted, results from systematizing.

MICHAEL D. RESNIK

University of North Carolina/Chapel Hill

## NEW BOOKS

ABBINNETT, ROSS: *Truth and Social Science: From Hegel to Deconstruction.* Thousand Oaks, CA: Sage, 1998. 200 p. Paper $24.95.

ACKERMAN, BRUCE: *We the People, Volume II: Transformations.* Cambridge: Harvard, 1998. xi, 515 p. Cloth $29.95.

ADAMS, WILLIAM Y.: *The Philosophical Roots of Anthropology.* Stanford: CSLI, 1998. xiii, 466 p. Cloth $54.95, paper $19.95.

ALMEDER, ROBERT: *Harmless Naturalism: The Limits of Science and the Nature of Philosophy.* Chicago: Open Court, 1998. xii, 235 p. Cloth $46.95, paper $18.95.

ANNAS, JULIA: *Platonic Ethics: Old and New.* Ithaca: Cornell, 1999. viii, 196 p. Cloth $35.00.

ARCHARD, DAVID: *Sexual Consent.* Boulder: Westview, 1998. x, 189 p. Cloth $55.00, paper $20.00.

ARDEN, JOHN BOGHOSIAN: *Science, Theology, and Consciousness.* Westport, CT: Praeger, 1998. xiv, 189 p. Cloth $55.00.

ARENDT, HANNAH: *The Human Condition, Second Edition.* Chicago: University Press, 1998. xx, 349 p. Cloth $45.00, paper $16.00.

AUDI, ROBERT: *Epistemology: A Contemporary Introduction to the Theory of Knowledge.* New York: Routledge, 1998. xii, 340 p. Cloth $70.00, paper $19.99.

AUSTIN, JOHN: *The Province of Jurisprudence Determined and the Uses of the Study of Jurisprudence.* Indianapolis: Hackett, 1998. xxxi, 396 p. Cloth $37.95, paper $16.95.

BALAGUER, MARK: *Platonism and Anti-Platonism in Mathematics* New York: Oxford, 1998. x, 217 p. Cloth $45.00.

BALKIN, J.M.: *Cultural Software: A Theory of Ideology.* New Haven: Yale, 1998. xii, 335 p. Cloth $35.00.

BARNETT, RANDY. *The Structure of Liberty: Justice and the Rule of Law.* New

York: Oxford, 1998. xi, 347 p. Cloth $29.95.

BARON, MARCIA W., PHILIP PETTIT, and MICHAEL STONE: *Three Methods of Ethics.* Malden, MA: Blackwell, 1997. 285 p. Paper $24.95.

BARTSCH, RENATE: *Dynamic Conceptual Semantics: A Logico-Philosophical Investigation into Concept Formation and Understanding.* Stanford: CSLI, 1998. vi, 296 p. Cloth $64.95, paper $24.95.

BECKER, LAWRENCE: *A New Stoicism.* Princeton: University Press, 1998. viii, 216 p. Cloth $29.95.

BELL, JEFFREY A.: *The Problem of Difference: Phenomenology and Poststructuralism.* Toronto: University Press, 1998. ix, 294 p. Cloth $75.00, paper $24.95.

BELLIOTTI, RAYMOND ANGELO: *Stalking Nietzsche.* Westport, CT: Greenwood, 1998. 187 p. Cloth $55.00.

BENJAMIN, ANDREW: *Present Hope: Philosophy, Architecture, Judaism.* New York: Routledge, 1998. xii, 180 p. Cloth $75.00, paper $19.99.

BERNSTEIN, MARK H.: *On Moral Considerability: An Essay on Who Morally Matters.* New York, Oxford, 1998. 189 p. Cloth $35.00.

BERNSTEIN, RICHARD J.: *Freud and the Legacy of Moses.* New York: Cambridge, 1998. xii, 151 p. Cloth $59.95, paper $18.95.

BICKLE, JOHN: *Psychoneural Reduction: The New Wave.* Cambridge: MIT, 1998. xi, 240 p. Cloth $35.00.

BLAIS, ANDREW L.: *On the Plurality of Actual Words.* Amherst: Massachusetts UP, 1997. 232 p. Cloth $35.00.

BLATTNER, WILLIAM D.:*Heidegger's Temporal Idealism.* New York: Cambridge, 1999. xvii, 325 p. Cloth $54.95.

BLOOR, DAVID: *Wittgenstein, Rules and Institutions.* New York: Routledge, 1997. xvi, 173 p. Cloth $65.00.

BOBIK, JOSEPH: *Aquinas on Matter and Form and the Elements: A Translation and Interpretation of the* De Principiis Naturae *and the* De Mixtione Elementorum *of St. Thomas Aquinas.* Notre Dame: University Press, 1998. xviii, 325 p. Cloth $39.00, paper $19.00.

BOK, HILARY: *Freedom and Responsibility.* Princeton: University Press, 1998. 220 p. Cloth $39.50.

BONJOUR, LAURENCE: *In Defense of Pure Reason.* New York: Cambridge, 1998. xiv, 232 p. Cloth $54.95, paper $18.95.

BOWIE, NORMAN E.: *Business Ethics: A Kantian Perspective.* Malden, MA: Blackwell, 1999. viii, 184 p. Paper $24.95.

BRAYBROOKE, DAVID: *Moral Objectives, Rules, and the Forms of Social Change.* Toronto: University Press, 1998. xvi, 364 p. Cloth $65.00, paper $21.95.

BRECKMAN, WARREN: *Marx, the Young Hegelians, and the Origins of Radical Social Theory.* New York: Cambridge, 1999. xii, 335 p. Cloth $54.95.

BRENT, JOSEPH: *Charles Sanders Peirce: A Life, Revised Edition.* Bloomington: Indiana UP, 1998. xx, 412 p. Paper $18.95.

BRINKLEY, ALAN: *Liberalism and Its Discontents.* Cambridge: Harvard, 1998. xii, 372 p. Cloth $27.95.

BRUDNEY, DANIEL: *Marx's Attempt to Leave Philosophy* Cambridge: Harvard, 1998. xviii, 425 p. Cloth $45.00.

BUNZL, MARTIN. *Real History: Reflections on Historical Practice.* New York: Routledge, 1998. viii, 152 p. Cloth $65.00, paper $19.99.

BUTCHVAROV, PANAYOT: *Skepticism about the External World.* New York: Oxford, 1998. 184 p. Cloth $45.00.

CAMPBELL, RICHMOND: *Illusions of Paradox. A Feminist Epistemology Naturalized.* Lanham, MD: Rowman and Littlefield, 1998. xii, 245 p. Cloth $58.00, paper $24.95.

CAMPBELL, SUE: *Interpreting the Personal: Expression and the Formation of Feelings.* Ithaca: Cornell, 1997. x, 204 p. Cloth $39.95, paper $15.95.

CARLSON, THOMAS A.: *Indiscretion: Finitude and the Naming of God.* Chicago: University Press, 1999. xiii, 307 p. Cloth $45.00, paper $18.00.

CARPENTER, BOB: *Type-Logical Semantics.* Cambridge: MIT, 1997. xxi, 575 p. Paper $35.00.

CARROLL, NOËL: *A Philosophy of Mass Art.* New York: Oxford, 1998. xii, 425 p. Cloth $75.00, paper $18.95.

CARTER, STEPHEN: *Civility: Manners, Morals, and the Etiquette of Democracy.* New York: Basic, 1998. xiv, 338 p. Cloth $25.00.

CAVE, ERIC M.: *Preferring Justice. Rationality, Self-Transformation, and the Sense of Justice.* Boulder: Westview, 1996. xiv, 183 p. Cloth $55.00.

CHAI, LEON: *Jonathan Edwards and the Limits of Enlightenment Philosophy.* New York: Oxford, 1998. xi, 164 p. Cloth $39.95.

CHURCHLAND, PAUL M., and PATRICIA S. CHURCHLAND: *On the Contrary: Critical Essays, 1987-1997.* Cambridge: MIT, 1998. xii, 349 p. Cloth $30.00.

CLACK, BRIAN R.: *Wittgenstein, Frazer, and Religion.* New York: St. Martin's, 1999. x, 200 p. Cloth $55.00.

CLEVELAND, TIMOTHY: *Trying without Willing: Essay in the Philosophy of Mind.* Brookfield, VT: Ashgate, 1997. xiii, 188 p. Cloth $59.95.

COATES, PETER: *Nature: Western Attitudes since Ancient Times.* Berkeley: California UP, 1998. viii, 246 p. Cloth $29.95.

COLLINS, ARTHUR: *Possible Experience: Understanding Kant's* Critique of Pure Reason. Berkeley: California UP, 1999. xix, 200 p. Cloth $40.00, paper $15.95.

CONWAY, DAVID: *Classical Liberalism: The Unvanquished Ideal, Second Edition.* New York: St. Martin's, 1998. ix, 150 p. Paper $21.95.

COOPER, JOHN M.: *Reason and Emotion: Essays on Ancient Moral Psychology and Ethical Theory.* Princeton: University Press, 1999. xvi, 588 p. Cloth $75.00, paper $26.95.

CORRINGTON, ROBERT C.: *Nature's Religion.* Lanham, MD: Rowman and Littlefield, 1997. xviii, 192 p. Cloth $58.00, paper $22.95.

COTTINGHAM, JOHN: *Philosophy and the Good Life: Reason and the Passions in Greek, Cartesian and Psychoanalytic Ethics.* New York: Cambridge, 1998. xiii, 230 p. Cloth $54.95, paper $17.95.

COWEN, TYLER: *In Praise of Commercial Culture.* Cambridge, Harvard, 1998. ix, 278 p Cloth $27.95.

CRISP, ROGER: *Mill on Utilitarianism.* New York· Routledge, 1997. xiv, 232 p. Cloth $60.00, paper $12.99.

CRITCHLEY, SIMON: *Very Little...Almost Nothing.* New York: Routledge, 1997. xi, 216 p. Cloth $75.00, paper $18.99.

CRITES, STEPHEN: *Dialectic and Gospel in the Development of Hegel's Thinking.* University Park: Penn State UP, 1998. xvii, 572 p. Cloth $65.00.

CUOMO, CHRIS J.: *Feminism and Ecological Communities.* New York: Routledge, 1998. x, 170 p. Cloth $65.00, paper $19.99.

CURD, PATRICIA: *The Legacy of Parmenides: Eleatic Monism and Later Presocratic Thought.* Princeton: University Press, 1998. x, 280 p. Cloth $45.00.

DADLEZ, E.M.: *What's Hecuba to Him? Fictional Events and Actual Emotions.* University Park: Penn State UP, 1997. viii, 240 p. Cloth $35.00, paper $16.95.

DALLMAYR, FRED: *Alternative Visions: Paths in the Global Village.* Lanham, MD: Rowman and Littlefield, 1998. xii, 307 p. Cloth $68.00, paper $24.95.

D'AMICO, ROBERT: *Contemporary Continental Philosophy.* Boulder: Westview, 1999. xi, 267 p. Cloth $59.00, paper $21.00.

DAVIS, MICHAEL: *Thinking Like an Engineer: Studies in the Ethics of a Profession.* New York: Oxford, 1998. xii, 240 p. Cloth $49.95.

DAVIS, STEPHEN T.: *God, Reason, and Theistic Proofs.* Grand Rapids, MI: Eerdmans, 1997. xiv, 204 p. Paper $26.00.

DAVIS, WAYNE A.: *Implicature: Intention, Convention, and Principle in the Failure of Gricean Theory.* New York: Cambridge, 1998. viii, 206 p. Cloth $49.95.

DENNETT, DANIEL C.: *Brainchildren: Essays on Designing Minds.* Cambridge: MIT, 1998. Paper $20.00.

DILMAN, ILHAM: *Love: Its Forms, Dimensions and Paradoxes.* New York: St. Martin's, 1998. xxi, 239 p. Cloth $59.95.

DOUBLE, RICHARD: *Beginning Philosophy.* New York: Oxford, 1999. xiii, 347 p. Paper $29.95.

DOVRING, FOLKE: *Knowledge and Ignorance: Essays on Lights and Shadows.* Westport, CT: Praeger, 1998. viii, 167 p. Cloth $55.00.

DURAN, JANE: *Philosophies of Science: Feminist Theories.* Boulder: Westview, 1998. xiv, 206 p. Cloth $60.00, paper $20.00.

EDMUNSON, WILLIAM A.: *Three Anarchical Fallacies: An Essay on Political Authority.* New York: Cambridge, 1998. xi, 192 p. Cloth $49.95.

ELDRIDGE, MICHAEL: *Transforming Experience: John Dewey's Cultural Instrumentalism.* Nashville: Vanderbilt, 1998. xii, 236 p. Cloth $45.00, paper $21.95.

ELDRIDGE, RICHARD: *Leading a Human Life: Wittgenstein, Intentionality, and Romanticism.* Chicago: University Press, 1997. xi, 300 p. Cloth $46.00, paper $17.95.

ELKINS, JAMES: *On Pictures and the Words That Fail Them.* New York: Cambridge, 1998. xix, 326 p. Cloth $65.00.

FEFERMAN, SOLOMON: *In the Light of Logic.* New York: Oxford, 1998. xii, 340 p. Cloth $49.95.

FERRE, FREDERICK: *Knowing and Value: Toward a Constructive Postmodern Epistemology*. Albany: SUNY, 1998. xviii, 393 p. Paper $29.95.

FESTENSTEIN, MATTHEW: *Pragmatism and Political Theory: From Dewey to Rorty*. Chicago: University Press, 1998. x, 237 p. Cloth $50.00, paper $19.00.

FIGAL, GUNTER: *For a Philosophy of Freedom and Strife. Politics, Aesthetics, Metaphysics*. Albany: SUNY, 1998. xi, 226 p. Cloth $59.50, paper $19.95.

FINNIS, JOHN: *Aquinas*. New York: Oxford, 1998. xxi, 385 p. Paper $18.95.

FLINT, THOMAS P.: *Divine Providence: The Molinist Account*. Ithaca: Cornell, 1998. xii, 258 p. Cloth $35.00.

FODOR, JERRY A.: *Concepts: Where Cognitive Science Went Wrong*. New York: Oxford, 1998. ix, 174 p. Cloth $55.00, paper $15.95; *In Critical Condition: Polemical Essays on Cognitive Science and the Philosophy of Mind*. Cambridge: MIT, 1998. x, 219 p. Cloth $29.50.

FORSTER, MICHAEL N.: *Hegel's Idea of a Phenomenology of Spirit*. Chicago: University Press, 1998. xi, 661 p. Cloth $75 00, paper $30.00.

FOTT, DAVID: *John Dewey. America's Philosopher of Democracy* Lanham, MD: Rowman and Littlefield, 1998. x, 167 p. Cloth $58.00, paper $22.95.

FRANKFURT, HARRY G.: *Necessity, Volition, and Love*. New York: Cambridge, 1999. xii, 180 p. Cloth $54.95, paper $17.95.

FRENCH, PETER A.: *Cowboy Metaphysics*. Lanham, MD: Rowman and Littlefield, 1997. xii, 162 p. Cloth $52.00, paper $21.95

GALLAGHER, SHAUN: *The Inordinance of Time*. Evanston, IL: Northwestern, 1998. xi, 243 p. Cloth $69.95, paper $22.95.

GALLOIS, ANDRÉ: *Occasions of Identity: A Study in the Metaphysics of Persistence, Change, and Sameness*. New York: Oxford, 1998. 296 p. Cloth $65.00.

GELLNER, ERNEST: *Language and Solitude: Wittgenstein, Malinowski, and the Habsburg Dilemma*. New York: Cambridge, 1998. xiii, 204 p. Cloth $54.95, paper $19.95.

GELVEN, MICHAEL: *The Risk of Being: What It Means to Be Good and Bad*. University Park: Penn State UP, 1997. viii, 169 p. Cloth $45.00.

GERT, BERNARD: *Morality. Its Nature and Justification*. New York: Oxford, 1998. xiv, 408 p. Paper $19.95.

GERT, BERNARD, CHARLES M. CULVER, and K. DANNER CLOUSER: *Bioethics: A Return to Fundamentals*. New York: Oxford, 1998. xii, 320 p. Cloth $34.95.

GEWIRTH, ALAN: *Self-Fulfillment*. Princeton: University Press, 1998. x, 235 p. Cloth $35.00.

GLAUSSER, WAYNE: *Locke and Blake: A Conversation across the Eighteenth Century*. Gainesville: Florida UP, 1998. xi, 201 p. Cloth $49.95.

GOEHR, LYDIA: *The Quest for Voice: On Music, Politics, and the Limits of Philosophy*. Berkeley: California UP, 1998. x, 237 p. Cloth $45.00.

GOLDBERG, SANFORD, and ANDREW PESSIN: *Gray Matters: An Introduction to the Philosophy of Mind*. Armonk, NY: M.E. Sharpe, 1997. xii, 304 p. Cloth $64.95, paper $23.95.

GONZALEZ, FRANCISCO J.: *Dialectic and Dialogue: Plato's Practice of Philosophical Inquiry*. Evanston, IL: Northwestern, 1998. x, 418 p. Cloth $89.95, paper $29.95.

GOVIER, TRUDY: *Dilemmas of Trust.* Montreal: McGill-Queen's, 1998. ix, 241 p. Cloth $29.95.

GRAHAM, GEORGE: *Philosophy of Mind: An Introduction, Second Edition.* Malden, MA: Blackwell, 1998. x, 267 p. Paper $24.95.

GRAHAM, GORDON: *Philosophy of the Arts: An Introduction to Aesthetics.* New York: Routledge, 1997. xi, 193 p. Cloth $65.00, paper $20.99.

GRAY, JOHN: *Hayek on Liberty.* New York: Routledge, 1998. xi, 187 p. Paper $24.99.

GRIFFIN, DAVID RAY: *Unsnarling the World-Knot: Consciousness, Freedom, and the Mind-Body Problem.* Berkeley: California UP, 1998. xv, 266 p. Cloth $45.00.

GRIM, PATRICK, GARY MAR, and PAUL ST. DENIS with the GROUP FOR LOGIC AND FORMAL SEMANTICS: *The Philosophical Computer: Exploratory Essays in Philosophical Computer Modeling.* Cambridge: MIT, 1998. viii, 322 p. Cloth $50.00.

GRISWOLD JR., CHARLES L.: *Adam Smith and the Virtues of the Enlightenment.* New York: Cambridge, 1999. xiv, 412 p. Cloth $59.95, paper $21.95.

GROH, FRED: *Due Respect: The Morality of the Welfare State.* Brookfield, VT: Ashgate, 1998. vi, 207 p. Cloth $63.95.

GUPTA, BINA: *The Disinterested Witness: A Fragment of Advaita Vedānta Phenomenology.* Evanston, IL: Northwestern, 1998. xx, 236 p. Cloth $59.95, paper $18.95.

HAACK, SUSAN: *Manifesto of a Passionate Moderate: Unfashionable Essays.* Chicago: University Press, 1998. x, 223 p. Cloth $22.95.

HACKING, IAN: *Mad Travelers: Reflections on the Reality of Transient Mental Illnesses.* Charlottesville: Virginia UP, 1998. x, 239 p. Cloth $27.95.

HAJI, ISHTIYAQUE: *Moral Appraisability: Puzzles, Proposals, and Perplexities.* New York: Oxford, 1998. xii, 272 p. Cloth $49.95.

HALLETT, GARTH L.: *Priorities and Christian Ethics.* New York: Cambridge, 1998. xiii, 202 p. Cloth $54.95.

HAMPTON, JEAN E.: *The Authority of Reason.* New York: Cambridge, 1998. xii, 310 p. Cloth $59.95, paper $18.95.

HANCOCK, CURTIS L., and BRENDAN SWEETMAN: *Truth and Religious Belief.* Armonk, NY: M.E. Sharpe, 1998. xii, 234 p. Cloth $54.95, paper $19.95.

HANSSEN, BEATRICE: *Walter Benjamin's Other History.* Berkeley: California UP, 1998. xi, 207 p. Cloth $35.00.

HARE, R.M.: *Sorting Out Ethics.* New York: Oxford, 1997. vii, 191 p. Cloth $18.95.

HARLE, VILHO: *Ideas of Social Order in the Ancient World.* Westport, CT: Greenwood, 1998. xvii, 252 p. Cloth $65.00.

HAUGELAND, JOHN: *Having Thought: Essays in the Metaphysics of Mind.* Cambridge: Harvard, 1998. x, 390 p. Cloth $29.95.

HAUSMAN, DANIEL M.: *Casual Asymmetries.* New York: Cambridge, 1998. xv, 300 p. Cloth $54.95.

HELM, PAUL: *Faith and Understanding.* Grand Rapids, MI: Eerdmans, 1997. viii, 212 p. Paper $26.00.

THE JOURNAL OF PHILOSOPHY
VOLUME XCVI, NO. 5, MAY 1999

UNIVERSITY LIBRARY
ALIPUR CAMPUS
3-1-2000

# RESPONSIBILITY, LUCK, AND CHANCE: REFLECTIONS ON FREE WILL AND INDETERMINISM*

Ludwig Wittgenstein[1] once said that "to solve the problems of philosophers, you have to think even more crazily than they do" (*ibid.*, p. 75). This task (which became even more difficult after Wittgenstein than it was before him) is certainly required for the venerable problem of free will and determinism.

## I. THE LUCK PRINCIPLE

Consider the following principle:

> (LP) If an action is *undetermined* at a time *t*, then its happening rather than not happening at *t* would be a matter of *chance* or *luck*, and so it could not be a *free* and *responsible* action.

This principle (which we may call the *luck principle*, or simply LP) is false, as I shall explain shortly. Yet it seems true. LP and a related principle to be considered later in this paper are fueled by many of those "intuition pumps," in Daniel Dennett's[2] apt expression, which

---

* This paper was prompted by a recent objection made in various forms against my view and other incompatibilist views of freedom and responsibility by Galen Strawson, Alfred Mele, Bernard Berofsky, Bruce Waller, Richard Double, Mark Bernstein, and Ishtiyaque Haji. (See footnote 10 for references.) The paper has benefitted from interchanges with the above persons and with participants at a conference on my work on free will at the University of Arkansas in September, 1997: Gary Watson, Barry Loewer, Timothy O'Connor, Randolph Clarke, Christopher Hill, and Thomas Senor. It has also benefitted from interchanges in conferences or in correspondence with John Martin Fischer, William Rowe, Nicholas Nathan, David Hodgson, Saul Smilansky, Kevin Magill, Peter van Inwagen, Derk Pereboom, Laura Ekstrom, Hugh McCann, and Ilya Prigogine. I am especially grateful to Mele and Strawson for pursuing me assiduously on these issues since the publication of my latest work, and for perceptive comments on the penultimate draft by Mele, Berofsky, and George Graham.

[1] *Culture and Value* (New York: Blackwell, 1980)

[2] *Elbow Room* (Cambridge: MIT, 1984), chapter 1 and pp. 32-34, 64-65, 119-20, 169-70

0022-362X/99/9605/217–40

support common intuitions about freedom and responsibility. LP and related principles lie behind the widespread belief that indeterminism, so far from being required for free will and responsibility, would actually undermine free will and responsibility. Dennett does not dwell on the intuition pumps of this sort, as I shall do in this paper. As a compatibilist, he is more interested in criticizing intuition pumps that lead people to think (mistakenly, on his view) that freedom and responsibility are not compatible with determinism, whereas intuition pumps that support LP lead people to think freedom and responsibility are not compatible with *indeterminism*. Yet intuition pumps of the latter kind are every bit as pervasive and influential in free-will debates as those Dennett dwells upon; and they are as much in need of deconstruction, since they play a significant role in leading people to believe that freedom and responsibility must be compatible with determinism.

I think the modern route to compatibilism—which is the reigning view among contemporary philosophers—usually goes through principles like LP at some point or other. In my experience, most ordinary persons start out as natural incompatibilists. They believe there is some kind of conflict between freedom and determinism; and the idea that freedom and responsibility might be compatible with determinism looks to them at first like a "quagmire of evasion" (William James) or "a wretched subterfuge" (Immanuel Kant). Ordinary persons have to be talked out of this natural incompatibilism by the clever arguments of philosophers—who, in the manner of their mentor, Socrates, are only too happy to oblige. To weaken natural incompatibilist instincts, philosophers first argue that what we mean by freedom in everyday life is the power or ability to do whatever we choose or desire to do—in short, an absence of coercion, compulsion, oppression, and other impediments or constraints upon our behavior. They then point out that we can be free in these everyday senses to do what we choose or desire, even if our choices and desires are determined by causes that lie in our past.

But this line of argument does not usually dispose of incompatibilist intuitions by itself. Ordinary persons might grant that many everyday freedoms are compatible with determinism and still wonder if there is not also some deeper freedom—the freedom to have an *ultimate* say in what we choose or desire to do in the first place—that is incompatible with determinism. (I have argued elsewhere[1] that this deeper freedom is what was traditionally meant by "free *will*.") So the philosophers must add a second step to their case—an argument to the effect that any allegedly deeper freedom (of the will) that is not compatible with determinism is no intelligible freedom at all. And with this step, principles like LP come into the picture. For any freedom not compatible with determinism would require indeterminism; and what is undetermined, it seems, would happen by chance or luck and could not be a free and responsible action. This kind of argument is the one that usually puts the final nail in the coffin of incompatibilist instincts.

---

[1] *The Significance of Free Will* (New York: Oxford, 1996), pp. 10-14, 33-37

When philosophy professors go through this two-stage argument in the modern classroom, they are replicating the standard case against traditional (incompatibilist or libertarian) free will which is one of the defining characteristics of modernity. The goal is to consign incompatibilist freedom to the dustbin of history with other beliefs that a modern scientific age is encouraged to outgrow. Students and ordinary persons subjected to this argument may have an uneasy feeling they are being had by the clever arguments of philosophers. But, also seeing no obvious response, except an appeal to mystery, many of them become compatibilists.

## II INDETERMINISM, THE BOGEYMAN

The second stage of this two-stage argument in support of compatibilism will concern me here, the one that goes through LP and related principles in the attempt to show that indeterminism would not enhance, but in fact would undermine, freedom and responsibility. What is at stake here is not merely the clever arguments of philosophers; for it happens that the case for principles like LP is a powerful one. It *is* difficult to see how indeterminism and chance can be reconciled with freedom and responsibility. Philosophers have tried to bring this out in a number of ways which will be addressed here. We may think of these as the varied intuition pumps that support LP and principles like it.

(1) We are often asked to consider, for example, that whatever is undetermined or happens by chance is not under the *control* of anything, and so is not under the control of the agent. But an action that is not under the control of the agent could not be a free and responsible action. (Here it is evident that the notion of control is involved in the case for LP: indeterminism and chance imply lack of control to a degree that implies lack of freedom and responsibility.)

(2) Another line of argument often heard is this: suppose a choice occurred as the result of an undetermined event (say, a quantum jump) in one's brain. Would that be a free choice? Being undetermined, it would appear to be more of a fluke or accident than a free and responsible action. Some twentieth-century scientists and

philosophers have suggested that free will might be rescued by supposing that undetermined quantum events in the brain could be amplified to have large-scale effects on choice or action.[4] Unfortunately, this modern version of the ancient Epicurean "swerve" of the atoms seems to be subject to the same criticisms as its ancient counterpart. It seems that undetermined events in the brain or body, whether amplified or not, would occur spontaneously and would be more of a nuisance—or perhaps a curse, like epilepsy—than an enhancement of freedom and responsibility.

(3) Nor would it help to suppose that the indeterminism or chance came *between* our choices (or intentions) and our actions. Imagine that you are intending to make a delicate cut in a fine piece of cloth, but because of an undetermined twitching in your arm, you make the wrong cut. Here, indeterminism is no enhancement of your freedom, but a *hindrance* or *obstacle* to your carrying out your purposes as intended. Critics of libertarian freedom[5] have often contended that this is what indeterminism would always be—a hindrance or impediment to one's freedom. It would get in the way, *diminishing* control, and hence responsibility, instead of enhancing them.

(4) Even more absurd consequences follow if we suppose that indeterminism or chance is involved in the initiation of overt actions. Arthur Schopenhauer[6] imagined the case of a man who suddenly found his legs start to move *by chance*, carrying him across the room against his wishes. Such caricatures are popular among critics of indeterminist freedom for obvious reasons: undetermined or chance-initiated overt actions would represent the opposite of controlled and responsible actions.

(5) Going a little deeper, one may also note that, if a choice or action is undetermined, it might occur otherwise *given exactly the same past and laws of nature* up to the moment when it does occurs. This means that, if Jane is deliberating about whether to vacation in Hawaii or Colorado, and gradually comes to favor and choose Hawaii, she might have chosen otherwise (chosen Colorado), given *exactly the same deliberation* up to the moment of choice that in fact led

---

[4] For example, physicist A. H. Compton, *The Freedom of Man* (New Haven: Yale, 1935) and neurophysiologist John Eccles, *Facing Reality* (New York: Springer, 1970).

[5] See, for example, Galen Strawson, who argues that, even if free will should be incompatible with determinism, indeterminism would be "no help" in enhancing either freedom or responsibility—"The Unhelpfulness of Indeterminism," *Philosophy and Phenomenological Research* (forthcoming)

[6] *Essay on the Freedom of the Will* (Indianapolis: Bobbs-Merrill, 1960), p. 47.

her to favor and choose Hawaii (exactly the same thoughts, reasonings, beliefs, desires, dispositions, and other characteristics—not a sliver of difference). It is difficult to make sense of this. The choice of Colorado in such circumstances would seem irrational and inexplicable, capricious and arbitrary.[7] If it came about by virtue of undetermined events in Jane's brain, this would not be an occasion for rejoicing in her freedom, but for consulting a neurologist about the waywardness of her neural processes.

(6) At this point, some defenders of incompatibilist freedom appeal to Gottfried Leibniz's[8] celebrated dictum that prior reasons or motives need not determine choice or action, they may merely "incline without necessitating"—that is, they may incline the agent toward one option without determining the choice of that option. This may indeed happen. But it will not solve the present problem; for it is precisely *because* Jane's prior reasons and motives (beliefs, desires, and the like) incline her toward the choice of Hawaii that choosing Colorado by chance at the end of exactly the same deliberation would be irrational and inexplicable. Similarly, if her reasons had inclined her toward Colorado, then choosing Hawaii by chance at the end of the same deliberation would have been irrational and inexplicable. And if prior reasons or motives had not inclined her either way (the celebrated medieval "liberty of indifference") and the choice was a matter of chance, then the choosing of one rather than the other would have been all the more a matter of luck and out of her control. (One can see why libertarian freedom has often been ridiculed as a mere "liberty of indifference.")

(7) Indeed, critics of indeterminist freedom have often argued that indeterminist free choices must always amount to *random* choices of this sort and hence the outcomes would be matters of mere luck or chance—like spinning a wheel to select among a set of alternatives. Perhaps there is a role for such random choices in our lives when we are genuinely indifferent to outcomes.[9] But to suppose that *all* of our free and responsible choices—including momentous ones, like whether to act heroically or treacherously—had to be by random selection in this way has been regarded by many philosophers as a *reductio ad absurdum* of the view that free will and responsibility require indeterminism.

---

[7] This dilemma for incompatibilist accounts of freedom is nicely described by Thomas Nagel, *The View from Nowhere* (New York: Oxford, 1986), chapter 7.

[8] *Selections* (New York: Scribner's, 1951), p. 435.

[9] Stephen M. Cahn makes a persuasive case for there being such a role—"Random Choices," *Philosophy and Phenomenological Research*, xxxvii (1977): 549-51.

(8) Consider one final argument which cuts more deeply than the others and to which I shall devote considerable attention. This paper was in fact prompted by new versions of this argument advanced in recent years against my incompatibilist account of free will by Galen Strawson, Alfred Mele, Bernard Berofsky, Bruce Waller, Richard Double, Mark Bernstein, and Ishtiyaque Haji[10]—though the argument is meant to apply generally to any view requiring that free actions be undetermined up to the moment when they occur.

Suppose two agents had exactly the same pasts (as indeterminism requires) up to the point where they were faced with a choice between distorting the truth for selfish gain or telling the truth at great personal cost. One agent lies and the other tells the truth. As Waller puts it, if the pasts of these two agents "are really identical" in every way up to the moment of choice, "and the difference in their acts results from chance," would there "be any grounds for distinguishing between [them], for saying that one deserves censure for a selfish decision and the other deserves praise" (*op. cit.*, p. 151)? Mele poses the problem in terms of a single agent in different possible worlds. Suppose in the actual world, John fails to resist the temptation to do what he thinks he ought to do, arrive at a meeting on time. If he could have done otherwise given the same past, then his counterpart, John* in a nearby possible world, which is the same as the actual world up to the moment of choice, resists the temptation and arrives on time. Mele then argues that, "if there is nothing about the agents' powers, capacities, states of mind, moral character and the like that explains this difference in outcome,...the difference is just a matter of luck" (*op. cit.*, pp. 582-83). It would seem that John* got lucky in his attempt to overcome temptation, whereas John did not. Would it be just to reward the one and punish the other for what appears to be ultimately the luck of the draw?

---

[10] Strawson, "The Impossibility of Moral Responsibility," *Philosophical Studies*, LXXV (1994): 5-24, and "The Unhelpfulness of Indeterminism," *op. cit.*; Mele, Review of my *The Significance of Free Will*, this JOURNAL, XCV, 11 (November 1998): 581-84, and "Luck and the Significance of Free Will," *Philosophical Explorations* (forthcoming); Berofsky, "Ultimate Responsibility in a Deterministic World," *Philosophy and Phenomenological Research* (forthcoming); Waller, "Free Will Gone Out of Control," *Behaviorism*, XVI (1988): 149-67; Double, *The Non-reality of Free Will* (New York: Oxford, 1991), p. 140; Bernstein, "Kanean Libertarianism," *Southwest Philosophy Review*, XI (1995): 151-57; Haji, "Indeterminism and Frankfurt-type Examples," *Philosophical Explorations* (forthcoming). Different, but related, concerns about indeterminism and agency are aired by Timothy O'Connor, "Indeterminism and Free Agency: Three Recent Views," *Philosophy and Phenomenological Research*, LIII (1993): 499-526; and Randolph Clarke, "Free Choice, Effort and Wanting More," *Philosophical Explorations* (forthcoming).

Considerations such as (1)-(8) lie behind familiar and varied charges that undetermined choices or actions would be "arbitrary," "capricious," "random," "uncontrolled," "irrational," "inexplicable," or "matters of luck or chance," and hence not free and responsible actions. These are the charges which principles like LP are meant to express. Responses to them in the history of philosophy have been many; but none to my mind has been entirely convincing. The charges have often led libertarians—those who believe in an incompatibilist free will—to posit "extra factors" in the form of unusual species of agency or causation (such as noumenal selves, immaterial egos, or nonoccurrent agent causes) to account for what would otherwise be arbitrary, uncontrolled, inexplicable, or mere luck or chance. I do not propose to appeal to any such extra factors in defense of libertarian freedom. Such appeals introduce additional problems of their own without, in my view, directly confronting the deep problems about indeterminism, chance, and luck to which considerations (1)-(8) are pointing. To confront these deep problems directly, I believe one has to rethink issues about indeterminism and responsibility from the ground up, without relying on appeals to extracausal factors—a task to which I now turn.

## III INDETERMINISM AND RESPONSIBILITY

First, one must question the intuitive connection in people's minds between "indeterminism's being involved in something's happening" and "its happening merely as a matter of chance or luck." 'Chance' and 'luck' are terms of ordinary language which carry the connotation of "its being out of my control" (as in (1) and (4) and above). So using them already begs certain questions, whereas 'indeterminism' is a technical term that merely precludes *deterministic* causation (though not causation altogether). Second, one must emphasize that indeterminism does not have to be involved in all free and responsible acts, even for incompatibilists or libertarians.[11] Frequently, we act from a will already formed; and it may well be that our actions are determined in such cases by our then existing characters and motives. On such occasions, to do otherwise by chance *would* be a fluke or accident, irrational and inexplicable, as critics of indeterminist freedom contend (in (3) and (4) above).

---

[11] I defend this point at length in *Free Will and Values* (Albany: SUNY, 1985), chapters 4 and 5. It is also defended by Peter van Inwagen, "When Is the Will Free?" in J. Tomberlin, ed., *Philosophical Perspectives*, Volume 3 (Atascadero, CA: Ridgeview, 1989), pp. 399-422. John Martin Fischer has described the view that van Inwagen and I defend as "restricted libertarianism," and has criticized it in "When the Will Is Free," in Tomberlin, ed., *Philosophical Perspectives*, Volume 6 (Atascadero: Ridgeview, 1992), pp. 423-51. Another critic is Hugh McCann, "On When the Will Is Free," in G. Holmstrom-Hintikka and R. Tuomela, eds., *Contemporary Action Theory*, Volume 1 (Dordrecht: Kluwer, 1997), pp. 219-32. van Inwagen responds to Fischer in "When Is the Will Not Free?" *Philosophical Studies*, LXXV (1994): 95-114; and I respond in *The Significance of Free Will*, pp. 32-43.

Incompatibilists about free will should not deny this. What they should rather say is that when we act from a will already formed (as we frequently do), it is "our own free will" by virtue of the fact that we formed it (at least in part) by earlier choices or actions which were not determined and for which we could have done otherwise voluntarily, not merely as a fluke or accident. I call these earlier undetermined actions *self-forming actions* or SFAs.[12] Undetermined SFAs are a subset of all of the actions done of our own free wills (many of which may be determined by our earlier formed character and motives). But if there were no such undetermined SFAs in our lifetimes, there would have been nothing we could have ever voluntarily done to make ourselves different than we are—a condition that I think is inconsistent with our having the kind of responsibility for being what we are which genuine free will requires.

Now, let us look more closely at these undetermined SFAs. As I see it, they occur at times in life when we are torn between competing visions of what we should do or become. Perhaps we are torn between doing the moral thing or acting from self-interest, or between present desires and long-term goals, or we are faced with difficult tasks for which we have aversions. In all such cases, we are faced with competing motivations and have to make an effort to overcome temptation to do something else we also strongly want. In the light of this picture, I suggest the following incompatibilist account of SFAs.[13] There is a tension and uncertainty in our minds at such times of inner conflict which are reflected in appropriate regions of our brains by movement away from thermodynamic equilibrium—in short, a kind of stirring up of chaos in the brain that makes it sensitive to micro-indeterminacies at the neuronal level. As a result, the uncertainty and inner tension we feel at such soul-searching moments of self-formation is reflected in the indeterminacy of our neural processes themselves. What is experienced phenomenologically as uncertainty corresponds physically to the opening of a window of opportunity that temporarily screens off complete determination by the past. (By contrast, when we act from predominant motives or settled dispositions, the uncertainty or indeterminism is muted. If it were involved then, it *would* be a mere nuisance or fluke, capricious or arbitrary, as critics contend (in (2), (5) and (6) above).)

---

[12] See *The Significance of Free Will*, pp 74-78 SFAs are also sometimes called "self-forming willings" or SFWs in that work (pp. 125ff.).

[13] This, in broad outline, is the account developed in my *The Significance of Free Will*, chapters 8-10. In later sections below, I make important additions to it in response to criticisms.

When we do decide under such conditions of uncertainty, the outcome is not determined because of the preceding indeterminacy—and yet it can be willed (and hence rational and voluntary) either way owing to the fact that in such self-formation, the agents' prior wills are divided by conflicting motives. If we overcome temptation, it will be the result of our effort; and if we fail, it will be because we did not *allow* our effort to succeed. And this is owing to the fact that, while we wanted to overcome temptation, we also wanted to fail, for quite different and incommensurable reasons. When we decide in such circumstances, and the indeterminate efforts we are making become determinate choices, we *make* one set of competing reasons or motives prevail over the others then and there *by deciding*.

Return now to concerns about indeterminism and responsibility in the light of this picture. Consider a businesswoman who faces a conflict in her will of the kind typically involved in such SFAs. She is on the way to a meeting important to her career when she observes an assault in an alley. An inner struggle ensues between her moral conscience, to stop and call for help, and her career ambitions, which tell her she cannot miss this meeting—a struggle she eventually resolves by turning back to help the victim. Now suppose this woman visits some future neuroscientists the next day and they tell her a story about what was going on in her brain at the time she chose, not unlike the story just told. Prior to choice, there was some indeterminacy in her neural processes stirred up by the conflict in her will. The indeterminism made it uncertain (and undetermined) whether she would go back to help or press onward.

Suppose further that two recurrent and connected neural networks are involved in the neuroscientists' story. Such networks circulate impulses and information in feedback loops and generally play a role in complex cognitive processing in the brain of the kind that one would expect to be involved in human deliberation. Moreover, recurrent networks are nonlinear, thus allowing (as some recent research suggests) for the possibility of chaotic activity, which would contribute to the plasticity and flexibility human brains display in creative problem solving (of which practical deliberation is an example).[14] The input of one of these recurrent networks consists of the woman's moral motives, and its output the choice to go back; the input of the other, her career ambitions, and its output, the choice to go on to her meeting. The two networks are connected, so that the indeterminism that made it uncertain that she would do the moral thing was coming from her desire to do the opposite, and vice versa—the indeterminism thus arising, as we said, from a conflict in the will. When her effort to overcome self-interested desires succeeded, this corresponded to one of the neural pathways reaching an activation threshold, overcoming the indeterminism generated by the other.

To this picture, one might now pose the following objection: if it really was undetermined which choice the woman would make (in neural terms, which network would activate) right up to the moment when she chose, it seems that it would be a matter of luck or chance that one choice was made rather than the other, and so she could not be held responsible for the outcome. (Note that this is an expression of LP.) The first step in response is to recall a point made earlier: we must be wary of moving too hastily from 'indeterminism is involved in something's happening' to 'its happening merely as a matter of chance or luck'. 'Luck' and 'chance' have meanings in ordinary language that mere indeterminism may not have. The second step is to note that indeterminism of itself does not necessarily undermine control and responsibility.[15] Suppose you are trying to think through a difficult problem (say, a mathematical problem) and there is some indeterminacy in your neural processes complicating the task—a kind of chaotic background. It would be like trying to concentrate and solve a problem with background noise or distraction. Whether you are going to succeed in solving the mathematical problem is uncertain and undetermined because of the distracting neural noise. Yet if you concentrate and solve the problem nonetheless, I think we can say that you did it and are responsible for doing it even though it was undetermined whether you would succeed. The indeterministic noise would have been an obstacle to your solving the problem which you nevertheless overcame by your effort.

---

[14] See P Huberman and G. Hogg, "Phase Transitions in Artificial Intelligence Systems," *Artificial Intelligence*, XXXIII (1987): 155-72; C. Skarda and W. Freeman, "How Brains Make Chaos in Order to Make Sense of the World," *Behavior and Brain Sciences*, X (1987)· 161-95; A. Babloyantz and A. Destexhe, "Strange Attractors in the Human Cortex," in L. Rensing, ed , *Temporal Disorder in Human Oscillatory Systems* (New York: Springer, 1985), pp. 132-43.

[15] Important recent defenses of the claim that indeterminism does not necessarily undermine control and responsibility include Clarke, "Indeterminism and Control," *American Philosophical Quarterly*, XXXII (1995): 125-38; Carl Ginet, *On Action* (New York: Cambridge, 1990), chapter 6, O'Connor; and Laura Ekstrom, *Free Will* (Boulder: Westview, forthcoming).

There are numerous other examples in the philosophical literature of this kind, where indeterminism functions as an obstacle to success without precluding responsibility. Consider an assassin who is trying to kill the prime minister, but might miss because of some undetermined events in his nervous system which might lead to a jerking or wavering of this arm. If he does hit his target, can he be held responsible? The answer (as J. L. Austin and Philippa Foot[16] successfully argued decades ago) is "yes," because he intentionally and voluntarily succeeded in doing what he was *trying* to do—kill the prime minister. Yet his killing the prime minister was undetermined. We might even say in a sense that he got lucky in killing the prime minister, when he could have failed. But it does not follow, if he succeeds, that killing the prime minister was not his action, not something he did; nor does it follow, as LP would require, that he was not responsible for killing the prime minister. Indeed, if anything is clear, it is that he both killed the prime minister and was responsible for doing so.

Or consider a husband who, while arguing with his wife, swings his arm down in anger on her favorite glass table top, intending to break it. Again we suppose that some indeterminism in the husband's efferent neural pathways makes the momentum of his arm indeterminate, so it is undetermined if the table will break right up to the moment when it is struck. Whether the husband breaks the table or not is undetermined. Yet it does not follow, if he succeeds, that breaking the table was not something he did; nor again does it follow, as LP would require, that he was not responsible for breaking it.[17] The inference sanctioned by LP from 'it was undetermined' to 'he was not responsible', is not valid. The above cases are counterexamples to it; and there are many more.

## IV. POSSIBLE WORLDS AND LP*

But one may grant this and still object that counterexamples to LP of these kinds do not amount to genuine exercises of free will involving

---

[16] Austin, "Ifs and Cans," in his *Philosophical Papers* (New York: Oxford, 1961), pp. 153-80; Foot, "Free Will as Involving Determinism," in Berofsky, ed., *Free Will and Determinism* (New York: Harper and Row, 1966), pp. 95-108.

[17] We must, of course, assume in both these examples that other (compatibilist) conditions for responsibility are in place—for example, that, despite his anger, the husband was not acting compulsively and would have controlled himself, if he had wished; that he knew what he was doing and was doing it intentionally to anger his wife, and so on (and similarly for the assassin). But the point is that nothing in the facts of either case preclude these assumptions from also being satisfied.

SFAs, such as the businesswoman's, where there is conflict in the wills of the agents and they are supposed to choose freely and responsibly *whichever* way they choose. If the assassin and husband succeed in doing what they are trying to do (kill the prime minister, break the table) they will do it *voluntarily* (in accordance with their wills) and *intentionally* (knowingly and purposely). But if they *fail* because of the indeterminism, they will not fail voluntarily and intentionally, but "by mistake" or "accident," or merely "by chance." Thus, their "power" to do *otherwise* (if we should even call it a power) is not the usual power we associate with freedom of choice or action in self-formation, where the agents should be able to choose or act either way voluntarily or intentionally. The power to do otherwise of the assassin and the husband is more like Jane's "power" in (5) and (6) of section II, to choose to vacation in Colorado by a fluke or accident, after a long deliberation in which she had come to favor Hawaii.

As a consequence, while LP may fail for cases like those of the assassin, husband, and mathematical problem solver, another luck principle similar to LP might still be applicable to genuine exercises of free will involving SFAs, like the businesswoman's: if it is undetermined at *t* whether an agent *voluntarily* and *intentionally* does *A* at *t* or *voluntarily* and *intentionally* does otherwise, then the agent's doing one of these rather than the other at *t* would be a matter of *luck* or *chance*, and so could not be a free and responsible action. This principle—let us call it LP*—is fueled by the same intuitions that fuel LP. Indeed, it is a special case of LP, but one that is more difficult to deal with because it is not subject to counterexamples like those of the husband and the assassin; and it seems to be applicable to SFAs, like the businesswoman's, where failure is not merely a matter of mistake or accident.

To explore further the difficulties posed by LP*, let us look at the final and, I think, most powerful of the intuition pumps in support of LP-type principles mentioned in section II, namely, consideration (8). This was the argument of Strawson, Mele, Berofsky, Waller, Double, Bernstein, and Haji about two agents, or one agent in different possible worlds, with the same pasts.

Consider the version of this argument by Mele, which appeared in this JOURNAL and is a particularly revealing and challenging version of it. In the actual world, an agent John succumbs to the temptation to arrive late to a meeting, whereas his counterpart, John*, in a nearby possible world, whose physical and psychological history is the same as John's up to the moment of choice (as indeterminism requires), resists this temptation. Similarly, we can imagine a counterpart to the businesswoman, businesswoman*, in a nearby possible world who goes to her meeting instead of stopping to aid the assault victim, given the same past. But then, Mele argues, "if there is nothing about [these] agents' powers, capacities, states of mind, moral character and the like that explains this difference in outcome," since they are the same up to the moment of choice in the two possible worlds, "then the difference is just a matter of luck" (*op. cit.*, p. 583).[18] It would seem that John* got lucky in his attempt to overcome temptation, whereas John did not; and similarly, the businesswoman got lucky in her attempt to overcome temptation, while businesswoman* did not.

Let us first consider a general form of this argument that would support LP.

(a) In the actual world, person *P* (for example, John, the businesswoman) does *A* at *t*.

On the assumption that the act is undetermined at *t*, we may imagine that:

(b) In a nearby-possible world which is the same as the actual world up to *t*, *P** (*P*'s counterpart with the same past) does otherwise (does *B*) at *t*.
(c) But then (since their pasts are the same), there is nothing about the agents' powers, capacities, states of mind, characters, dispositions, motives, and so on prior to *t* which explains the difference in choices in the two possible worlds.
(d) It is therefore a matter of luck or chance that *P* does *A* and *P** does *B* at *t*.
(e) *P* is therefore not responsible (praiseworthy or blameworthy, as the case may be) for *A* at *t* (and presumably *P** is also not responsible for *B*).

Call this the *luck argument*. The key assumption is the assumption of indeterminism, which leads to step (b). The remaining steps are meant to follow from (b), given (a).

Despite the fact that this argument looks like Mele's and has an initial plausibility, it is not his argument—and it is a good thing it is not. For the argument from (a)-(e) is invalid as it stands—for the same reasons that LP was invalid. Consider the husband and husband* (his counterpart in a nearby world who fails to break the wife's table). If the outcome is undetermined, husband and husband* also have "the same powers, capacities, states of mind, characters, dispositions, motives, and so on" up to the moment of breaking or not breaking the table, as the argument requires; and it is a matter of luck or chance that the table breaks in one world and not the other. But for all that, it does not follow, as (e) requires, that the husband is not responsible for breaking the table. The husband would have quite a task persuading his wife that he was not responsible for breaking the table on the grounds that it was a matter of luck or chance that it broke. ("Luck or chance did it, not me" is an implausible excuse.)

---

[18] I have elsewhere denied that the pasts of the agents can be exactly the same, since, with indeterminist efforts, there is no exact sameness or difference (*The Significance of Free Will*, pp. 171-74). Mele's argument is designed to work, however, whether this denial of exact sameness is assumed or not. So I do not make an issue of it here.

But, of course, as we noted, husband* is not also responsible for *failing* to break the table, since he does not fail to break it voluntarily or intentionally. He is responsible only for the attempt, when he fails. Similarly, assassin* would be responsible for the attempted murder of the prime minister, when he missed. What has to be explicitly added to the argument (a)-(e) to avoid counterexamples like these is the LP* requirement that *both* $P$ and $P^*$ *voluntarily* and *intentionally* do $A$ and $B$ respectively in their respective worlds. Specifically, we must add to premise (a) that $P$ voluntarily and intentionally does $A$ at $t$ and to (b), that $P^*$ voluntarily and intentionally does $B$ at $t$, and then make the corresponding additions to (d) and (e). This will yield what we might call the LP* version of the luck argument rather than the LP version. And the stronger LP* version is clearly the one Mele intends, since John's choice in his example is supposed to be an SFA, like the businesswoman's choice in my example, where the agents can go either way voluntarily and intentionally. Moreover, this version of the argument—like LP* itself—is immune to counterexamples like those of the husband and the assassin.

### V PARALLEL PROCESSING

Nonetheless, despite immunity from these counterexamples, I think the LP* version of the luck argument, and LP* itself, also fail. But it is far less easy to show why. To do so, we have to take a closer look at SFAs and push the argument beyond where it has come thus far. Let it be granted that the businesswoman's case and other SFAs like John's are not like the examples of the husband and the assassin. The wills of the husband and assassin are already "set" on doing what they intend, whereas the wills of agents in SFAs, like the businesswoman and John, are not already settled or "formed" until they choose (hence the designation "self-forming actions").[19]

---

[19] See *The Significance of Free Will*, pp. 112-14.

Thus, to get from examples like those of the husband and assassin to genuine SFAs, I think we must do two things. First, we must put the indeterminacy involved in the efferent neural pathways of the husband and assassin into the central neural processes of the businesswoman and other agents, like John, who are making efforts of will to overcome moral, prudential, and other temptations. This move has already been made in earlier sections. But to respond to LP* versions of the luck argument, like Mele's, I believe this move must also be combined with another—a kind of "doubling" of the example given earlier of solving the mathematical problem in the presence of background indeterministic noise.[20]

Imagine that the businesswoman is *trying* or making an effort to solve *two* cognitive problems at once, or to complete two competing (deliberative) tasks at once—to make a moral choice and to make a choice for her ambitions (corresponding to the two competing neural networks involved in the earlier description). With respect to each task, as with the mathematical problem, she is being thwarted in her attempt to do what she is trying to do by indeterminism. But in her case, the indeterminism does not have a mere external source; it is coming from her own will, from her desire to do the opposite. Recall that the two crossing neural networks involved are connected, so that the indeterminism which is making it uncertain that she will do the moral thing is coming from her desire to do the opposite, and vice versa. She may therefore fail to do what she is trying to do, just like the assassin, the husband, and the person trying to solve the mathematical problem. But I argue that, if she nevertheless *succeeds*, then she can be held responsible because, like them, she will have succeeded in doing *what she was trying to do*. And the interesting thing is that this will be true of her, *whichever choice is made*, because she was trying to make both choices and one is going to succeed.

Does it make sense to talk about agents trying to do two competing things at once in this way? Well, we know the brain is a parallel processor and that capacity, I believe, is essential for the exercise of free will. In cases of self-formation, agents are simultaneously trying to resolve plural and competing cognitive tasks. They are, as we say, of two minds. But they are not therefore two separate persons. They are not disassociated from either task.[21] The businesswoman who wants to go back and help the assault victim is the same ambitious woman who wants to go on to her meeting and close the sale. She is a complex creature, like most of us who are often torn inside; but hers is the kind of complexity needed for free will. And when she succeeds in doing one of the things she is trying to do, she will endorse that as *her* resolution of the conflict in her will, voluntarily and intentionally, as LP* requires. She will not disassociate from either outcome, as did Jane (in (5) of section II), who wondered what "happened to" her when she chose Colorado, or like the husband and assassin who did not also want to fail.[22]

---

[20] This further "doubling" move is consistent with the theory put forward in *The Significance of Free Will*, and presupposes much of that theory, but is not made in that work. It is a further development especially provoked by Mele's argument discussed here as well as by criticisms of other persons since the book's publication, such as Strawson, Berofsky, Nicholas Nathan, Gary Watson, Clarke, O'Connor, Double, and Haji

But one may still object that the businesswoman makes one choice rather than the other *by chance*, since it was undetermined right up to the last moment which choice she would make. If this is so, we may have the picture of her first making an effort to overcome temptation (to go on to her meeting) and do the moral thing, and then at the last minute "chance takes over" and decides the issue for her. But this is the wrong picture. On the view just described, you cannot separate the indeterminism from the effort to overcome temptation in such a way that *first* the effort occurs *followed by* chance or luck (or vice versa). One must think of the effort and the indeterminism as fused; the effort *is* indeterminate and the indeterminism is a property of the effort, not something separate that occurs after or before the effort. The fact that the woman's effort of will has this property of being indeterminate does not make it any less her *effort*. The complex recurrent neural network that realizes the effort in the brain is circulating impulses in feedback loops and there is some indeterminacy in these circulating impulses. But the whole process is her effort of will and it persists right up to the moment when the choice is made. There is no point at which the effort stops and chance "takes over." She chooses *as a result of* the effort, even though she might have failed because of the indeterminism.

---

[21] I account for this elsewhere in terms of the notion of a "self-network" (*The Significance of Free Will*, pp 137-42), a more comprehensive network of neural connections representing the general motivational system in terms of which agents define themselves as agents and practical reasoners. For further discussion of such a notion, see Owen Flanagan, *Consciousness Reconsidered* (Cambridge: MIT, 1992), pp. 207ff.

[22] In response to my claim (*The Significance of Free Will*, p. 215) that "free willers [who engage in SFAs] are always trying to be better than they are by their own lights," by trying to overcome temptations of various sorts, Strawson asks: but "can't they also try to be worse than they are?"—"The Unhelpfulness of Indeterminism." He is right, of course; they can. I should have added what I am saying here, that free willers can and do *also* try to be as bad or worse than they are by resisting efforts to be better. Strange creatures indeed.

And just as expressions like 'She chose *by* chance' can mislead us in these contexts, so can expressions like 'She got lucky'. Ask yourself this question: Why does the inference 'He got lucky, *so he was not responsible*' *fail* when it does fail, as in the cases of the husband and the assassin? The first part of an answer goes back to the claim that 'luck', like 'chance', has question-begging implications in ordinary language which are not necessarily implications of "indeterminism" (which implies only the absence of deterministic causation). The core meaning of 'He got lucky', which *is* implied by indeterminism, I suggest, is that 'He succeeded *despite the probability or chance of failure*'; and this core meaning does not imply lack of responsibility, if he succeeds.

If 'He got lucky' had further meanings in these contexts often associated with 'luck' and 'chance' in ordinary usage (for example, the outcome was not his doing, or occurred by *mere* chance, or he was not responsible for it), the inference would not fail for the husband and assassin, as it clearly does. But the point is that these further meanings of 'luck' and 'chance' do not follow *from the mere presence of indeterminism*. Second, the inference 'He got lucky, so he was not responsible' fails because *what* the assassin and husband succeeded in doing was what they were trying and wanting to do all along. Third, *when* they succeeded, their reaction was not "Oh dear, that was a mistake, an accident—something that *happened* to me, not something I *did*." Rather, they *endorsed* the outcomes as something they were trying and wanting to do all along, that is to say, knowingly and purposefully, not by mistake or accident.

But these conditions are satisfied in the businesswoman's case as well, *either way* she chooses. If she succeeds in choosing to return to help the victim (or in choosing to go on to her meeting) (i) she will have "succeeded despite the probability or chance of failure"; (ii) she will have succeeded in doing what she was trying and wanting to do all along (she wanted both outcomes very much, but for different reasons, and was trying to make those reasons prevail in both cases); and (iii) when she succeeded (in choosing to return to help) her reaction was not "Oh dear, that was a mistake, an accident—something that happened to me, not something I did." Rather, she endorsed the outcome as something she was trying and wanting to do all along; she recognized it as her resolution of the conflict in her will. And if she had chosen to go on to her meeting she would have endorsed that outcome, recognizing it as her resolution of the conflict in her will.

## VI THE LUCK ARGUMENT REVISITED

With this in mind, let us return to the LP* version of the argument from (a)-(e). I said that Mele clearly intends this stronger LP* version of the argument, since the force of his argument depends on the fact that John's choice in his example is a SFA, like the businesswoman's, instead of being like the actions of the husband and assassin. But if this is so, then John's situation will also be like the businesswoman's on the account just given of SFAs. Since both of them are simultaneously trying to do *both* of the things they may do (choose to help or go on, overcome the temptation to arrive late or not), they will do either with intent or on purpose, as a result of wanting and trying to do it—that is, intentionally and voluntarily. Thus, their "failing" to do one of the options will not be a mistake or accident, but a voluntary and intentional doing *of the other.*

Likewise, businesswoman* and John* are simultaneously trying to do both things in their respective worlds; and they will not "fail" to act on moral or weak-willed motives by mistake or accident, as the case may be, but by voluntarily and intentionally choosing to act on the opposing motives. The point is that in self-formation of these kinds (SFAs), failing is never *just* failing; it is always also a *succeeding* in doing something else we wanted and were trying to do. And we found that one can be responsible for succeeding in doing what one was trying to do, even in the presence of indeterminism. So even if we add the LP* requirement of more-than-one-way voluntariness and intentionality to the argument of (a)-(e), the argument remains invalid for cases like the businesswoman's and other SFAs, like John's.

But one might argue further, as Mele does, that John and John* (and businesswoman and businesswoman*) not only had the same capacities, motives, characters, and the like prior to choice, but they made exactly the same *efforts* as well. And this does seem to suggest that the success of one and failure of the other was a matter of mere luck or chance, so that John and the businesswoman were not responsible. But again the inference is too hasty. Note, first, that husband and husband* also made the same efforts (as well as having the same capacities, motives, and characters) up to the very moment of breaking of the table. Yet it does not follow that the husband is not responsible when he succeeds. And *both* the businesswoman and businesswoman*, and John and John*, are in the position of the husband in their respective worlds, since both will have succeeded in doing what they were trying to do.

But one may still want to object: if the businesswoman and businesswoman*, and John and John*, make exactly the same efforts, how can it *not* be a matter of chance that one succeeds and the other does not, in a way that makes them not responsible? To which I reply: But if they both succeeded in doing what they were trying to do (because they were simultaneously trying to do both things), and then having succeeded, they both *endorsed* the outcomes of their respective efforts (that is, their choices) as what they were trying to do, instead of disowning or disassociating from those choices, how then can we *not* hold them responsible? It just does not follow that, because they made exactly the same efforts, they chose *by* chance.

To say something was done "by chance" usually means (as in the assassin and husband cases when they fail), it was done "by mistake" or "accidentally," "inadvertently," "involuntarily," or "as an unintended fluke." But none of these things holds of the businesswoman and John either way they choose. Unlike husband*, businesswoman* and John do not fail to overcome temptation by mistake or accident, inadvertently or involuntarily. They consciously and willingly fail to overcome temptation *by* consciously and willingly choosing to act in selfish or weak-willed ways. So, just as it would have been a poor excuse for the husband to say to his wife when the table broke that "Luck or chance did it, not me," it would be a poor excuse for businesswoman* and John to say "Luck or chance did it, not me" when they failed to help the assault victim or failed to arrive on time.

Worth highlighting in this argument is the point that we cannot simply say the businesswoman and businesswoman* (or John and John*) made exactly the same *effort* (in the singular) in their respective possible worlds and one succeeded while the other failed. We must say they made exactly the same *efforts* (plural) in their respective worlds. Mentioning only one effort prejudices the case, for it suggests that the failure of that effort in one of the worlds was a *mere* mistake or accident, when the fact is that both of the agents (P and P*) made *both* efforts in *both* worlds. In one world, one of the efforts issued in a choice and in the other world, a different effort issued in a different choice; but neither was merely accidental or inadvertent in either world. I would go even further and say that we may also doubt that the efforts they were both making really were exactly the same. Where events are indeterminate, as are the efforts they were making, there is no such thing as exact sameness or difference of events in different possible worlds. Their efforts were not exactly the same, nor were they exactly different, because they were not exact. They were simply unique.[23]

One might try another line: perhaps we are begging the question in assuming that the outcomes of the efforts of the businesswoman and her counterpart were *choices* at all. If they were not choices to begin with, they could not have been voluntary choices. One might argue this on the grounds that (A) "If an event is undetermined, it must be something that merely happens and cannot be somebody's choice"; and (B) "If an event is undetermined, it must be something that merely happens, it cannot be something an agent does (it cannot be an action)." But to see how question-begging these assumptions are, one has only to note that (A) and (B) imply respectively (A′) "If an event is a choice, it must be determined" ("All choices are determined") and (B′) "If an event is an action, it must be determined" ("All actions are determined"). Are these supposed to be a priori or analytic truths? If so, then long-standing issues about freedom and determinism would be settled by fiat. If an event were not determined, it could not be a choice or action necessarily or by definition.[24]

This explains the businesswoman's suspicions when she exited the neuroscientists' offices. They told her that when she "chose" to go back to help the assault victim the day before, there was some indeterminism in her neural processes prior to choice. She accepted this as a correct empirical finding. But she was suspicious when the neuroscientists tried to get her to make the further inference from those findings which she did not really *choose* to help the assault victim yesterday. She refused to accept that conclusion, and rightly so. For in drawing it, they were going beyond their empirical findings and trying to foist on her the a priori assumption that if an event was undetermined, it could not have been her choice or could not have been something she did. She rightly saw that there was nothing in the empirical evidence that required her to say that. To choose is consciously and deliberately to form an intention to do something; and she did that, despite the indeterminism in her neural processes (as did businesswoman* when she chose to go on to her meeting).

## VI. FINAL CONSIDERATIONS: CONTROL AND EXPLANATION

But it is one thing to say that she chose and another to say she chose *freely* and *responsibly*. This would require that she not only chose, but had voluntary *control* over her choice either way. We have not talked at length to this point about the matter of control (considerations (1) and (3) of section II) and must now do so. For this may be the reason why we may think the choices made by the businesswoman and businesswoman* (or John and John*) could not be responsible, if they were undetermined. We might deny that they had voluntary control over what they chose, where voluntary control means being able to bring about something in accordance with one's will or purposes (or, as we often say, the ability to bring something about "at will").

---

[23] See *The Significance of Free Will*, pp. 171-74.

[24] *Ibid.*, pp. 183-86, for a fuller account of why indeterminism does not rule out action or choice.

One thing does seem to be true about control which critics of indeterminist freedom have always maintained: indeterminism, wherever it appears, does seem to *diminish* rather than enhance agents' voluntary control (consideration (3) of section II). The assassin's voluntary control over whether or not the prime minister is killed (his ability to realize his purpose or what he is trying to do) is diminished by the undetermined impulses in his arm—and so also for the husband and his breaking the table. Moreover, this limitation is connected to another, which I think we must also grant—that indeterminism, wherever it occurs, functions as a *hindrance* or *obstacle* to our purposes that must be overcome by effort (consideration (3)).

But recall that in the businesswoman's case (and for SFAs generally, like John's), the indeterminism that is admittedly diminishing her ability to overcome selfish temptation, and *is* indeed a hindrance to her doing so, is coming from her own will—from her desire and effort to do the opposite—since she is simultaneously trying to realize two conflicting purposes at once. Similarly, her ability to overcome moral qualms is diminished by the fact that she also simultaneously wants and is trying to act on moral reasons. If we could look at each of the two competing neural networks involved separately, abstracting from the other, the situation would look analogous to the situations of the husband and the assassin. The agent would be trying to do something while being hindered by indeterminism coming from an external source. But, in fact, we cannot look at the two networks separately in this way because, in reality, they are connected and interacting. The indeterminism that is a hindrance to her fulfilling one is coming from its interactions with the other. The indeterminism, therefore, does not have an external source. It is internal to her will, and hence to her self, since she identifies with both networks and will identify with the choice reached by either of them as her choice.

The upshot is that, despite the businesswoman's diminished control over *each* option considered separately, due to a conflict in her will, she nonetheless has what I call *plural voluntary control* over the two options considered *as a set* (*ibid.*, pp. 134-43). Having plural voluntary control over a set of options means being able to bring about *whichever* of the options you will or most want, *when* you will to do so, for the reasons you will to do so, without being coerced or compelled in doing so. And the businesswoman (or John) has this power, because whichever of the options she chooses (to help the victim or go on to her meeting) will be *endorsed* by her as what she wills or most wants to do at the moment when she chooses it (though not necessarily beforehand); she will choose it for the reasons she most wants to act on then and there (moral or selfish reasons, as the case may be); she need not have been coerced by anyone else into choosing one rather than the other; and she will not be choosing either compulsively, since neither choice is such that she could not have chosen it then and there, even if she most wanted to.[23]

One must add, of course, that such plural voluntary control is not the same as what may be called *antecedent determining control*—the ability to determine or guarantee which of a set of options will occur *before* it occurs (*ibid.*, p. 144). With respect to undetermined self-forming choices (SFAs), agents cannot determine or guarantee which choice outcome will occur *beforehand*; for that could only be done by predetermining the outcome. But it does not follow that, because one cannot determine which of a set of outcomes will occur before it occurs, one does not determine which of them occurs *when* it occurs. When the conditions of plural voluntary control are satisfied, agents exercise control over their present and future lives then and there by deciding.

But can we not at least say that, if indeterminism is involved, then *which* option is chosen is "arbitrary"? I grant that there is a sense in which this is true. An ultimate arbitrariness remains in all undetermined SFAs because there cannot, in principle, be sufficient or overriding *prior* reasons for making one set of competing reasons prevail over the other. But I argue that such arbitrariness relative to prior reasons tells us something important about free will. It tells us, as I have elsewhere expressed it, that every undetermined self-forming choice (SFA) "is the initiation of a 'value experiment' whose justification lies in the *future* and is not fully explained by the *past*. [Making such a choice], we say in effect, 'Let's try this. It is not required by my past, but is consistent with my past and is one branching pathway my life could now meaningfully take. I am willing to take responsibility for it one way or the other' " (*ibid.*, pp. 145-46). To initiate and take responsibility for such value experiments whose justification lies in the future, is to "take chances" without prior guarantees of success. Genuine self-formation requires this sort of risk-taking and indeterminism is a part of it. If there are persons who need to be certain in advance just exactly what is the best or right thing to do in every circumstance (perhaps to be told so by some human or divine authority), then free will is not for them.

---

[23] *Ibid.*, pp. 133-38, where a more detailed case is made for each of these claims.

This point also throws light on why the luck argument fails, even in the stronger LP* version, despite its initial plausibility. Consider the move from step (c)—the agents $P$ and $P^*$ have the same powers, characters, motives, and the like, prior to $t$ in the two possible worlds—to step (d), which says it was a matter of luck or chance that $P$ did $A$ and $P^*$ did $B$ at $t$. An important reason given for this move was that, if both agents have all the same prior powers, characters, motives, and the like, there can be no "explanation of the difference in choice" between the two agents in terms of their prior reasons or motives; and this is taken to imply that the difference in choices in the two worlds is a matter of luck or chance *in a way* that precludes responsibility.

But this move, like others discussed earlier, is too hasty. The absence of an explanation of the difference in choice in terms of prior reasons does not have the tight connection to issues of responsibility one might initially credit it with. For one thing, the absence of such an explanation does not imply (as I have been arguing throughout this paper) that businesswoman and businesswoman* (John and John*) (1) did not *choose* at all, nor does it imply that they did not both choose (2) *as a result of their efforts*, nor that they did not choose (3) *for reasons* (different reasons, of course) that (4) they most wanted to choose for *when* they chose, nor that they did not choose for those reasons (5) *knowingly* and (6) *on purpose* when they chose, and hence (7) *rationally*, (8) *voluntarily*, and (9) *intentionally*. None of these conditions is precluded by the absence of an explanation of the difference of choice in terms of prior reasons. Yet these are precisely the kinds of conditions we look for when deciding whether or not persons are responsible.

I suggest that the reason why these conditions are not excluded is that the explanation of the difference of choice in the two possible worlds which is missing is an explanation in terms of *sufficient* or *conclusive* reasons—one that would render an alternative choice, given the same prior reasons, irrational or inexplicable. And, of course, *that* sort of explanation is not possible for undetermined SFAs, when there is conflict in the will and the agent has good (but not decisive or conclusive) prior reasons for going either way. But neither is that sort of explanation required to say that an agent acts as the result of her effort for reasons she most wants to act on then and there. In sum, *you can choose responsibly for prior reasons that were not conclusive or decisive prior to your choosing for them.*

I said a moment ago that such arbitrariness relative to prior reasons tells us something important about free will—that every self-forming choice is the initiation of a value experiment whose justification lies in the future and cannot be fully explained by the past. It is worth adding in this regard that the term 'arbitrary' comes from the Latin *arbitrium*, which means 'judgment'—as in *liberum arbitrium voluntatis* ("free judgment of the will")—the medieval designation for free will. Imagine a writer in the middle of a novel. The novel's heroine faces a crisis and the writer has not yet developed her character in sufficient detail to say exactly how she will react. The author must make a "judgment" (*arbitrium*) about how she will react that is not determined by the heroine's already formed past, which does not give unique direction. In this sense, the author's judgment of how she will act is "arbitrary," but not entirely so. It has input from the heroine's fictional past and, in turn, gives input to her projected future.

In a similar manner, agents who exercise free will are both authors of, and characters in, their own stories at once. By virtue of "self-forming" judgments of the will (*arbitria voluntatis*), they are "arbiters" of their own lives, taking responsibility for "making themselves" out of past that, if they are truly free, does not limit their future pathways to one. If someone should charge them with not having a sufficient or conclusive prior reason for choosing as they did, they may reply as follows: "Perhaps so. But that does not mean I did not *choose*; and it does not mean I did not choose for *good* reasons, which I stand by and for which I take responsibility. If I lacked sufficient or conclusive prior reasons, that is because, like the heroine of the novel, I was not a fully formed person before I chose—and still am not, for that matter.[38] Like the author of the novel, I am in the process of writing a story and forming a person (who, in my case, is myself). It is a heavy burden, but an eminently human one."

ROBERT KANE

University of Texas/Austin

---

[38] Jan Branson (in "Alternatives of Oneself," *Philosophy and Phenomenological Research* (forthcoming)) has made an important distinction that is relevant here—between choosing "alternatives *for* oneself" and choosing "alternatives *of* oneself" Branson notes that some choices in life are for different courses of action that will make a difference in what sort of person the chooser will become in future. In such cases, agents are not merely choosing alternatives for themselves but are choosing alternatives of themselves. Many SFAs, as I understand them, would be of this kind.

# INTENTION AND WEAKNESS OF WILL*

There is something curious about the philosophical literature on weakness of will. It is not about what one might expect. Even David Wiggins,[1] in a discussion that has much in common with that to be given here, starts out by claiming:

> Almost anyone not under the influence of theory will say that, when a person is weak-willed, he intentionally chooses that which he knows or believes to be the worse course of action when he could choose the better course (*ibid.*, p. 239).

I do not agree that this is the untutored view. Whenever I have asked nonphilosophers what they take weakness of will to consist in, they have made no mention of judgments about the better or worse course of action. Rather, they have said things like this: weak-willed people are irresolute; they do not persist in their intentions; they are too easily deflected from the path that they have chosen. My aim here is to pursue this line of thought. I shall develop the idea that the central cases of weakness of will are best characterized not as cases in which people act against their better judgment, but as cases in which they fail to act on their intentions. Of course, to say this is not to say enough. Not every case of a failure to act on one's intentions is a case of weakness of will. Sometimes, we realize that our intentions were ill-judged, or that circumstances have changed to make them inappropriate. When, in such cases, we fail to act on our intentions, we need not display weakness of will. So a large part of the discussion below will be concerned with saying what kind of failure to act on one's intention does constitute weakness of will. The basic idea is this. Just as it takes skill to form and employ beliefs, so it takes skill to form and employ intentions. In particular, if intentions are to fulfil their function, they need to be relatively resistant to reconsideration. Weakness of will arises, I shall suggest, when agents are too ready to reconsider their intentions.

---

* This paper was written while I was a fellow at the Research School of Social Sciences, Australian National University. Thanks to audiences there, and at Monash University, Sydney University, and Sheffield University for discussion; and especially to Sam Guttenplan, Lloyd Humberstone, Rae Langton, David Lewis, Philip Pettit, and Michael Smith.

[1] "Weakness of Will, Commensurability, and the Objects of Deliberation and Desire," in his *Needs, Values, Truth* (Cambridge: Basil Blackwell, 1987), pp. 239-67.

0022-362X/99/9605/241–62

In taking this approach, I depart from almost all of the literature on the subject.[2] Since I find the approach obvious, this calls for some explanation. I suspect that there are two factors at work. First, there has been, until quite recently, a widespread suspicion of the notion of intention. Some have said that intentions are a philosopher's invention, a misleading reification of the legitimate talk of intentional action. Others, while prepared to concede a place to intentions, have held them to be reducible to beliefs and desires, or else to beliefs and desires and their relations to action. Such considerations have led many philosophers to think that an account of weakness of will must ultimately be couched in the language of beliefs and desires. The traditional understanding of weakness of will as failure to do what one judges to be best is couched in this language; so, too, are the main rival accounts.[3] In the last few years, however, a number of philosophers have argued that talk of intentions is both legitimate and irreducible: I have in mind especially the work of Michael Bratman and Gilbert Harman.[4] If they are right that such talk is legitimate, then there is no *requirement* to analyze weakness of will in terms of belief and desire rather than in terms of intention. That is all that is needed for the account offered here. If, in addition, they are right that talk of intentions is irreducible, then it will not be possible to rephrase the account offered here in terms of beliefs and desires. This, however, is an issue that I shall not broach; although I am sympathetic to the irreducibility claim, I shall not presuppose it in what follows.

---

[2] Exceptions: Wiggins (*ibid.*) goes on to sketch an account of weakness of will that does tie it to failure to persist in one's intentions. Alfred Mele—*Irrationality* (New York: Oxford, 1987)—discusses cases of weakness of will in which one acts against one's intention, but he still treats this as a special (and specially problematic) case of action against one's best judgment; that is, these are cases in which one judges a particular course as right, and then forms the intention to pursue it, and then fails to act on it. In contrast, I suggest that action against one's judgments about what is right need not enter into examples of weakness of will at all. Amelie Rorty insists in various articles that weakness of will should not be identified with *akrasia*; she also charts a number of places at which the "akratic break" can take place, including at the point between forming an intention and acting on it—"Where Does the Akratic Break Take Place?" *Australasian Journal of Philosophy*, LVIII (1980): 333-46. I discuss some of the issues raised by her paper below Christopher Peacocke argues that an account of intention is essential to an account of *akrasia*—"Intention and *Akrasia*," in Bruce Vermazen and Merrill B. Hintikka, eds., *Essays on Davidson: Action and Events* (New York. Oxford, 1985), pp. 51-73; but his project is very different from that offered here. Roughly, he is concerned with the idea that *akratic* action results from *forming an intention* (instead of simply performing an action) that is not in line with one's judgment as to what is best.

[3] Namely, accounts that analyze weakness of will in terms of a failure to bring one's lower-order desires into line with one's higher order desires, as presented in John Bigelow, Susan Dodds, and Robert Pargetter, "Temptation and the Will," *American Philosophical Quarterly*, XXVII (1990): 39-49, following Richard C. Jeffrey, "Preference among Preferences," this JOURNAL, LXXI, 13 (July 18, 1974): 377-91; and accounts that analyze weakness of will in terms of desire revision that does not result from conditionalization, as presented in Frank Jackson, "Weakness of Will," *Mind*, XCIII (1984): 1-18.

[4] Bratman, *Intention, Plans, and Practical Reason* (Cambridge: Harvard, 1987); Harman, *Change in View* (Cambridge: MIT, 1986).

The second factor which has influenced the lines of the discussion is the impetus given by Plato and Aristotle. Both clearly were concerned with the question of whether it is possible to act against one's best judgment; and much of the contemporary literature has been concerned with assessing and responding to their arguments. Let us grant the term *'akrasia'* to refer to action voluntarily undertaken against one's best judgment; it is, after all, scarcely a term of ordinary English. Then my contention will be that weakness of will is not *akrasia*.

The outline of my discussion is as follows. In section I, I give an account of intentions, relying largely on Bratman's work, but going beyond him in some areas. I show in section II how this leads naturally to an account of weakness of will. In section III, I chart the considerable advantages that this account has over the traditional account that identifies weakness of will with *akrasia*. I reply to some possible objections in section IV.

## I INTENTIONS

Bratman's account of intentions has three parts. He identifies the characteristics that a state needs to have if we are to count it as an intention; he explains the useful role that states with these characteristics would have in our lives; and he gives reasons for thinking that beliefs and desires on their own could not fulfill this role. Taken together, these provide the basis for the case that we should acknowledge the existence of intentions. For if we seem to have them, and if having them would play a useful role, and if indeed they do seem to play this useful role, then the burden of proof rests in their favor. We shall need compelling arguments before giving them up.

I shall not be concerned here with the third element of Bratman's argument: the contention that beliefs and desires cannot play the role allotted to intentions. Readers in search of that can read Bratman's book. But the other two elements of his argument—the account of the distinguishing marks of intentions, and of the useful role that intentions can play—will be central to my account of weakness of will. Let me start by identifying the distinguishing marks; in particular, let me start with the distinguishing marks of *future-directed* intentions (that is, intentions to do something at some future time). I shall return to consider present-directed intentions later.

A future-directed intention to perform a certain action typically has the following characteristics. The agent forms the intention at one time by making a decision to perform the action. Then, unless it is revised, the intention will directly lead the agent to perform the action; it is, as Bratman says, *controlling* (*op. cit.*, p. 16). Moreover, it is relatively immune to reconsideration and hence to revision. Once formed, intentions have a tendency to persist. They have what Bratman calls *stability* (*op. cit.* pp. 16, 65).

What need do we have of states that are stable and controlling in this way? Part of Bratman's answer stems from the observation that we are epistemically limited creatures. Information is scarce and costly to obtain. Reasoning on the information that we have takes time and effort. It is rational then to limit the amount of time we spend looking for information and reasoning on it. We should do a certain amount of searching and reasoning, and then stop and make a decision on the basis of that. Now, if we were to act immediately on making the decision, then this would not in itself give rise to a need for intentions. But suppose our decision is a decision to act tomorrow. Then we need some way of storing our decision so that we act on it tomorrow without reconsidering it (for to reconsider it would violate the requirement that we not reason about it further). In short, we need a state that is stable and controlling: an intention to perform the action that is relatively immune to revision, and will in itself be enough to move us when the time comes.

This, in turn, raises another question: Why should we want to make a decision now about how we shall act tomorrow? It could be that now is a more propitious moment to reason. Tomorrow, there will be some factor that will stop us from reasoning well: we shall be short of time, or distracted, or some such. But a more common reason for doing our reasoning ahead of time is that all sorts of other actions will be dependent upon what we decide to do; and we shall need to perform some of these actions in the meanwhile. Thus, suppose I want to paint my front door tomorrow. Then there might be many reasons for deciding today what color to paint it. It could be that tomorrow I shall start at the crack of dawn when there will be too little light to see the color charts properly; that will give me a reason of the first kind for doing my deliberation today. More probably, the paint shop will be closed tomorrow, so I should buy the paint today; and since I want only to buy the color that I shall use, I shall need to have decided today what color that will be. I need to form my intention ahead of time. Thus, intentions are necessary in the *intrapersonal* case for agents to coordinate their activities. They are also important in the *interpersonal* case. Suppose some friends are coming to visit me tomorrow, and the easiest way for them to recognize my house is from the color of the front door. I shall be speaking with them today, but not again before they visit. Then, once again, I have a reason for forming today an intention about which color I shall paint my door tomorrow, so that I can tell them what to look for.

It will be important to my argument to realize that there can be good reason for forming intentions now, even in cases where I cannot make my decision just on the basis of rational deliberation. One such case involves Buridan-style examples: if I am really indifferent between red and green for my door, it is nonetheless important that I decide on one of them, and form the corresponding intention. Otherwise, I shall not know what paint to buy, or what to tell my visitors. A more interesting case concerns options that the agent finds incommensurable. Suppose (taking an example at random) that I do not know how to compare the demands of leaving home to fight fascism with the demands of staying to look after my mother. Here, the problem is not that each outcome fares equally well on the same scale; I do not know how to place them on the same scale. Nevertheless, I need to decide on one course, and to form the corresponding intention; both intrapersonal and interpersonal coordination require this. I need to know whether to start winding up my affairs and packing my bag; and I need to know what to tell my mother so that she too can make her plans accordingly.[5]

I have sketched the main lines of Bratman's argument for the utility of intentions. Once we have got this far, it is easy to see that they might be useful in other roles, too, roles that Bratman does not discuss. Suppose that there is an act that I now believe I should perform at some time in the future; and suppose I know that when the time comes, I shall not want to perform it. Then it would be useful to form an intention now, an intention which will lead me directly to act when the time comes, and which will be somewhat immune to reconsideration in the light of the desires I shall have then. Philosophers have been keen on the idea that sometimes an intention will not be enough; that we will need to bind ourselves in some further way to fulfil our intentions, perhaps by placing temptation out of our reach, or by telling others of our plans so that fear of their disapproval should we fall short will provide an extra incentive.[6] But we should not let our interest in these exotic methods blind us to the fact that very often, intention is enough. People get up on cold dark mornings, leave enjoyable lunches to return to work, even give up smoking, using no mechanism other than an intention to overcome their contrary desires.[7]

---

[5] Bratman points out that cases like these show the utility of intentions even in cases where information is not limited; for we have reason to think that Buridan cases will arise even for agents who know everything. He also argues, I think convincingly, that they show that we cannot reduce talk of intentions to talk of beliefs and desires. By hypothesis, I do not desire that I take one course rather than the other; and my intention to take one course cannot be seen as a belief that I will take it. See Bratman, "Davidson's Theory of Intention," in *Essays on Davidson: Actions and Events*, pp. 13-26, here pp. 22-23; *Intention, Plans, and Practical Reason*, pp. 11, 22ff.

The use of intention to overcome present desire becomes especially important in cases that involve repeated actions. I reason that since smoking forty cigarettes a day for the rest of my life will make a considerable difference to my chance of getting lung cancer, I should give it up. But should I deny myself the cigarette I was about to have? Smoking one cigarette will make virtually no difference to my chances of getting lung cancer. So why should I deny myself? Unfortunately, the same argument will work just as well forty times a day for the rest of my life. What we need here is not just an intention to perform—or, in this case, to refrain from performing—a specific action. We need a general intention concerning a certain type of action; what Bratman calls a *policy*,[8] and I shall call a *policy intention*.

To conclude this section, let us return to an issue shelved at its beginning: Are there such things as *present-directed* intentions? All the considerations raised so far have concerned future-directed intentions, that is, intentions to do something in the future. Do they give us any reason for accepting intentions to perform actions straight away? I rather doubt that they do; however, there is a slight theoretical elegance in accepting present-directed intentions. I spoke of future-directed intentions as resulting from a decision to act. I do not mean by this that there are two distinct states: the decision and the intention. Rather, to form the intention just is to make the decision.

---

[6] Jon Elster, *Ulysses and the Sirens* (New York: Cambridge, 1979). I suspect that very often the point of telling others of our intentions is not to incur their scorn if we revise them. Rather, it is simply to remind us of what our intentions were, or of how seriously we made them. A midwife told me of the entry on one woman's birth plan: "If I ask for pain relief in labor, show me this birth plan." The woman did ask for pain relief, the midwife showed her the birth plan, and the woman decided against taking any.

[7] Bratman briefly mentions examples in which intentions are used to overcome future resistance, but argues that we get a better understanding of the central cases of intention by ignoring them—*Intention, Plans, and Practical Reason*, p. 12. I suspect that such examples do provide central cases of intention; what we do well to ignore are examples in which intention is not enough.

[8] *Intention, Plans, and Practical Reason*, pp. 87-91.

(This is not to say that you cannot *have* an intention without having made a decision. You might have been born with the intention; or it might have been implanted by some manipulator using fiendish methods. But if so, you did not *form* the intention.) Then we might want to say that a decision to act now also results in an intention, but a present-directed one. A number of theorists have taken this path.[9] It seems to me to be slightly at odds with ordinary usage, but innocuously so; and it will make my account simpler if I adopt it here. So when I speak of intentions, these can be either present- or future-directed.

## II. WEAKNESS OF WILL

I have argued that we have a need for intentions, that is, that we have a need for action-guiding states that are not readily revised. My suggestion now is that this gives the basis for an account of weakness of will. A person exhibits weakness of will when she revises her intention, in circumstances in which she should not have revised it. This 'should' is not meant in a moral sense. Rather, it is the 'should' that is generated by the norms of the skill of managing one's intentions. A person is weak willed if she revises her intentions too readily.

What is it to be too ready to revise an intention? Clearly, not every case of a revision is a case of an over-ready revision. If you and I intend to go for a picnic, and it starts to rain, we might be quite right to revise our intention, and in doing so we need not show weakness of will. (I only say we 'need not' show weakness of will in such a case, for I have left it underdescribed; if we had vowed to go whatever the weather, then we probably would have been weak willed in giving up when it rained.) It might be thought that we can accommodate cases like these by invoking conditional intentions. Even if we did not say so explicitly, it might be thought that our intention was the conditional one of going for a picnic provided it did not rain. But not every reasonable revision of an intention can be understood as the triggering of an escape clause in a conditional intention. We would be equally right to abandon our intention in the face of a freak plague of frogs falling from the sky; but neither of us considered that, even implicitly. We need, then, some general account of when it is right to revise our intentions.

Friedrich Nietzsche[10] seems to have held that it is *always* wrong to revise an intention:

---

[9] See, for instance, the notion of proximal intentions in Mele, *Springs of Action* (New York: Oxford, 1992), part II.

[10] *Beyond Good and Evil* (Harmondsworth: Penguin, 1973); cited also in Gary Watson, "Skepticism about Weakness of Will," *Philosophical Review*, LXXXVI (1977): 316-39, here p. 328.

> To close your ears to even the best arguments once the decision has been made: sign of a strong character. Thus an occasional will to stupidity (*ibid.*, §107).

Perhaps this is good advice for football referees. But as a general attitude it is surely a recipe for disaster. Moreover, it is not even true that such intransigence is a sign of a strong character. One is reminded of Somerset Maugham's[11] jibe against the Vicar of Blackstable: "Like all weak men he laid an exaggerated stress on not changing one's mind" (*ibid.*, chapter 39). Nevertheless, if our account of intentions is correct, there is something right about the idea that there should be a real reluctance to revise an intention. To see this, let us return once more to Bratman, and to his account of the circumstances in which we rationally should reconsider our intentions. His idea is that it is rational to reconsider an intention just in case doing so manifests tendencies that it is reasonable for the agent to have; similarly, it is rational to fail to reconsider an intention just in case this manifests tendencies that it is reasonable for the agent to have.[12] This means, of course, that sometimes it will be rational not to reconsider an intention even when reconsideration would, in that instance, have been in the agent's interests. In general, once we have decided in which restaurant to eat, it is a good idea to let the matter rest, without endlessly discussing the pros and cons; and this is true, even though it occasionally means that we shall go to a restaurant that we could have realized was not the best choice.

Presumably, the converse of Bratman's criterion also holds: it is not rational to reconsider something if the reconsideration manifests tendencies that it is not reasonable for the agent to have. Adding this to the account proposed here, we arrive at the following: actors show weakness of will when they revise an intention as a result of a reconsideration that they should not have performed; that is, when their reconsideration exhibits tendencies that it is not reasonable for the agent to have.

---

[11] *Of Human Bondage* (New York: Doran, 1915). Note that Maugham does not ascribe weakness of will to the vicar. There is more than one kind of weakness; what is at issue here is more like weakness of character. For a further distinction, consider J. L. Austin's complaint about the "grotesque confusion of moral weakness with weakness of will"—"A Plea for Excuses," in his *Philosophical Papers* (New York: Oxford, 1961), pp. 123-52, here p. 146, note 1.

[12] Bratman, *Intention, Plans, and Practical Reason*, p. 68. He actually talks in terms of habits rather than tendencies. But for something to be a habit, it must be habitually exhibited. In contrast, it is possible for a tendency to be only rarely exhibited, if at all. I want the latter notion.

As it stands, this will seem horribly vague. Which tendencies of reconsideration is it reasonable for an agent to have? I would suggest rules of thumb like the following:

(1) It is reasonable to have a tendency to reconsider intentions, if one believes that circumstances have changed in such a way that they defeat the purpose of having the intention.
(2) It is reasonable to have a tendency to reconsider intentions, if one believes that they can no longer be carried out.[13]
(3) It is reasonable to have a tendency to reconsider intentions, if one believes that they will lead one to great suffering when that suffering was not envisaged at the time of forming the intention.
(4) It is reasonable to have a tendency not to reconsider intentions in circumstances that prevent clear thought, if those intentions were made in circumstances that allow clear thought.
(5) It is reasonable to have a tendency not to reconsider intentions that were expressly made in order to get over one's later reluctance to act.

This is still pretty vague: the principles themselves are vague, they are doubtless incomplete, and I have said nothing about which of them should take precedence over which. But I shall not try to eliminate this vagueness. Vagueness is a problem when we try to determine the extension of a concept: when we try to provide a means of telling for each particular act whether it does or does not display weakness of will. My concern here is rather with giving an analysis of the concept itself. If the concept is vague, then the analysis had better be vague, too, and along just the same dimensions. I hope that we shall see that the account proposed here provides just the right sort of vagueness.

The account is not quite plausible as it stands. I need to make one clarification and one amendment. First, the clarification. Suppose I intend to perform a rash and dangerous act: leaping from a high cliff on an untested homemade hang-glider. Would I show weakness of will in revising my intention? That rather depends on why I revise it. If I reassess the chances that the hang-glider will fail, and decide on a program of more cautious preliminary testing, that is surely a reasonable revision. If, on the other hand, I simply suffer a failure of nerve, backing off from the edge of the cliff at the sight of the drop before me, a failure I would have suffered no matter how good the hang-glider, then that is plausibly a case of weakness of will. So the question is not simply whether or not it is reasonable to reconsider my intention; it is whether or not the particular reconsideration I perform is one that exhibits tendencies that it is reasonable for me to have.

---

[13] Perhaps this rule is otiose. Perhaps it is a necessary condition on the *formation* of an intention that one does not believe that one will be unsuccessful in carrying it out. Or, more controversially, perhaps one must actually believe that one will be successful. (For discussion and references, see Bratman, *Intention, Plans, and Practical Reason*, pp. 35–41; he denies that these are necessary conditions.) Equally then, perhaps it is impossible to *maintain* an intention in the face of a new belief that one cannot carry it out, or even in face of the loss of the belief that one can.

Second, the amendment.[14] Some people go in for a lot of low-level intention revision. They decide to go to one restaurant, change their minds in favor of another, switch back to the first, again to the second, and so on. Now, it is not reasonable to have such tendencies; much time and effort is wasted, and interpersonal coordination becomes a nightmare. But while we might describe such people as fickle or capricious, we would not normally describe them as weak willed. So the account offered so far is too broad. We need to know what is distinctive about weakness of will, what it is that distinguishes it from caprice.

I mentioned before that one reason for forming an intention can be in an attempt to overcome contrary desires that one believes one will have when the time comes to act. An intention can be, as we might say, *contrary inclination defeating*. Now, this might be only part of the reason for forming a particular intention, and not a very important part at that. Indeed, sometimes this aspect might be merely implicit in an intention, and might only emerge if we asked the person what they would do if they were to change their preferences. The notion is not vacuous, however, since some intentions clearly are not contrary inclination defeating. Thus, someone might form an intention that has the content: I will go to The Red Lion tonight, providing I still feel like going (and such an intention can still have an important role in coordination, despite its conditional nature).

So we have a distinction between intentions which are designed to be contrary inclination defeating, and those which are not. This provides us with the grounds that we need for distinguishing weakness of will from caprice. If someone overreadily revises an intention that is, at least partially, contrary inclination defeating, that is weakness of will; if they overreadily revise an intention that is not, that is caprice. Consider again the vacillating diner. Suppose he has become concerned about his tendency to keep changing his mind, and so resolves to go to a particular restaurant even if another seems more attractive later on. In other words, suppose he forms a contrary-inclination-defeating intention to go to that restaurant. Then if he revises his intention once again, he would not merely be capricious; he would display weakness of will.

---

[14] Jeanette Kennett showed me that I needed to make an amendment here. Lloyd Humberstone showed me how to make it.

Thus, we can think in terms of a genus with two species. The genus is irresoluteness: unreasonable revision of intention. The two species, weakness of will and caprice, are distinguished by whether the intentions revised are contrary-inclination-defeating intentions or not. I am not sure how theoretically important this distinction is; moreover, we have now introduced further vagueness at the boundary between the two species. But the distinction, and the vagueness at the boundary, does, I think, correspond to our ordinary usage. Since this is a paper on weakness of will, I shall limit discussion to intentions that are contrary inclination defeating; but since that is rather a mouthful, I shall just call them *intentions*.

## III. ADVANTAGES OF THE PRESENT PROPOSAL

I shall present six reasons for believing the account of weakness of will which is offered here; or, to be more precise, six reasons for preferring it to the traditional account that understands weakness of will as a failure to do what one judges to be best.[15]

*(i) Accommodating weakness of will in cases of indifference or incommensurability*. I mentioned before that we form intentions in cases where we are indifferent between the alternatives, or treat them as incommensurable. When we do so, and revise our intentions, we can show weakness of will. That is just what the account proposed here would predict. But the traditional account can make no sense of such cases; for, by hypothesis, the agents involved do not judge one option to be better than the other, and so cannot be acting against their better judgment.

To see the possibility of such cases, consider again the example of incommensurability mentioned above. I am caught between the conflicting demands of fighting fascism and of staying to look after my mother. Unable to compare them in any meaningful way yet knowing I must choose, I decide to go and fight. But, then, when the moment to leave comes, I revise my intention and stay with my mother. That can be an example of weakness of will. (Note that to generate such a case we need not say that the options really are incommensurable; only that the agent judges them to be so.)

*(ii) Explaining the relation of weakness of will to strength of will.* One would expect the property opposed to weakness of will to be

---

[15] Some of these considerations also, I believe, provide reasons for preferring it to the accounts offered by Bigelow, Dodds, and Pargetter, and by Jackson; but I shall not argue for this contention here

strength of will. And strength of will is something that shades off into stubbornness. The account I propose explains this in a straightforward way. A person shows strength of will when they stick by their intentions in circumstances in which they are right to do so; that is, when they do not reconsider them, and not doing so exhibits tendencies that it is reasonable for them to have. Strength of will turns to stubbornness when they stick by their intentions even when it is reasonable to reconsider and revise them. And, of course, the boundary between these two is vague; reasonable people will disagree about which tendencies of intention revision it is reasonable to have, just as they will disagree on when strength of will moves over into stubbornness. (One way of understanding the quotation from Nietzsche is to see him placing the boundary at one extreme.) Here, then, is one place where the vagueness in the *analysans* mirrors the vagueness in the *analysandum*.

In contrast, the traditional account has real problems in explaining strength of will. Indeed, defenders of the traditional account have typically contrasted weakness of will with self-control, where this is understood as the ability to do what one believes to be best. But self-control is not continuous with stubbornness, in the way that strength of will is. It is perhaps possible to have too much self-control; but the worry there is that one lacks spontaneity; and lacking spontaneity is very different from being stubborn.

(*iii*) *Accounting for cases of oscillating weakness of will.* Consider this example from Thomas Schelling[16]:

> As a boy I saw a movie about Admiral Byrd's Antarctic expedition and was impressed that as a boy he had gone outdoors in shirtsleeves to toughen himself up against the cold. I resolved to go to bed at night with one blanket too few. That decision to go to bed minus one blanket was made by a warm boy. Another boy awoke cold in the night, too cold to retrieve the blanket and resolving to restore it tomorrow. But the next bedtime it was the warm boy again, dreaming of Antarctica, who got to make the decision. And he always did it again (*ibid.*, p. 59).

Let us suppose that the cold Schelling had at least tried to retrieve the blanket. How would we describe the situation? Would we say that Schelling displayed weakness of will when he reached for the blanket despite his earlier resolution to leave it off? Or would we say that he displayed weakness of will in leaving the blanket off again the following night, despite his resolution to leave it on? I think that we might well want to say both (we would need to fill out the details in the right way; he must form each intention seriously, having given it proper consideration, and so on). The account I am offering enables us to do so. We say that in each case, in overturning his previous intention, Schelling displays tendencies that it is not reasonable for him to have. But on the traditional account of weakness of will, it is not easy to see how we can get this result.

---

[16] "The Intimate Contest for Self-Command," in his *Choice and Consequence* (Cambridge: Harvard, 1984), pp. 57-82.

On the traditional account, the normal case is one in which the agent maintains a view about which action is the right one, and then acts contrary to this belief. But that cannot be the case here, since if the young Schelling maintained a view, then it would have been either that it is best to leave the blanket on, or that it is best to take it off. But if he maintained either of these views, then it is not possible for *both* actions to display weakness of will, since it is not possible for both actions to be contrary to the view maintained. Alternatively, proponents of the traditional account might think that the young Schelling changed his mind about which course was the best. Then, in order to characterize him as doubly weak willed, they would have to say that in the middle of the night, shivering beneath one blanket too few, he believed that what was best was to keep the blanket off, and his weakness stemmed from reaching for it; and they would have to say that when going to bed, warm and dreaming of the Antarctic, he believed that it was best to keep the blanket on, and his weakness stemmed from his leaving it off. But that is completely implausible; if there were any reason for thinking that, in the middle of the night, he believed it best to keep the blanket off, that would be because he believed this when he went to bed, and had not changed his mind; and similarly, if there were any reason for thinking that, at bedtime, he believed it best to keep the blanket on, that would be because he believed this in the middle of the night, and had not changed his mind.

*(iv) Allocating the stigma of weakness of will.* There is a considerable stigma attached to being weak willed. I think that the account offered here correctly predicts the times that it is deserved, and the traditional account does not. I have a friend who believes that all the arguments point to the same conclusion: he should not eat meat. But he is not moved. "I am," he says disarmingly, "inconsistent." Now there is something very odd about that, but, although he is *akratic*, he does not attract the stigma that attaches to weakness of will. Indeed, I would never call this friend weak willed. Suppose, however, that he announces that he has decided to give up meat; his intention is firm. If I then find him roasting himself a suckling pig in a pit, I

shall say that he is weak willed, and quite rightly. It is the failure to persist in the intention that makes all the difference.

We can make the point in another way. Suppose my friend tells me that he will give up eating meat on the first day of January; it is his New Year's resolution. I find him ruefully eating sausages on New Year's Eve; I cannot scorn him for his weakness of will. Yet were I to find him eating the same meal the following day, I could. His views about what is best have not changed over the two days; the difference stems from the intention.

These considerations are relevant to one of the binding strategies that was mentioned before. If I fear that I shall not persist in my intention, then one thing I can do is to make it public; for the scorn that I shall suffer if I revise the intention will provide me with an incentive to keep with it. The important point here is that, if I want to enlist the potential scorn of others in this way, it is not much use simply announcing my opinion of what is best; I need to announce my intention.

*(v) Explaining cases of weakness of will without inner conflict.* Some people say that in cases of weakness of will there must be an inner conflict: the agent must feel the pull of the course that they are weakly abandoning. And they take this as an argument for the traditional account.[17] That strikes me as quite wrong. Of course, when we self-ascribe weakness of will, we shall typically feel some tension; otherwise, we would not know to self-ascribe it. But we can ascribe weakness of will to a third party in whom there is no conflict. We surely can ascribe weakness of will to a person who has vowed to give up smoking, and who blithely starts up again straight away, saying that they have changed their mind (again, we would need to fill in the details in the right way). We can make perfect sense of this on the view proposed here: we simply say that they gave up their intention too easily. We need to be sure that the initial intention really was formed, and too sudden a change might make us doubt that; and the revision itself might be seen as involving a kind of conflict. But there need be no internal conflict at all at the point at which they smoke the cigarette.

Thus it seems that rather than be a point in the traditional account's favor, its requirement that there be internal conflict actually counts against it. For it is hard to see that the traditional account could count our recidivist smoker as weak willed unless there was, deep down and repressed, a belief that smoking was the wrong thing to do. Yet we do not need to attribute such self-deception in order to attribute weakness of will.

---

[17] For instance, Christopher Cordner, "Jackson on Weakness of Will," *Mind*, XCIV (1985): 273-80

Note that I have not said that cases of *akrasia* are impossible. Suppose that I were tempted by the arguments of Plato and Aristotle and their latter-day followers, arguments to the conclusion that a person cannot choose other than that which they judge best. Then I might try to exploit the account given here to say that what appears to be *akrasia* is really just an unreasonable revision of intention. But I am not tempted by those arguments. So I can happily accept that there are cases of *akrasia*, and that they do give rise to internal conflict. I simply claim that unless they also involve an unreasonable revision of intention, they are not cases of weakness of will.

*(vi) Accommodating cases of both* akrasia *without weakness of will, and weakness of will without* akrasia. So far we have considered cases of weakness of will without *akrasia*, and we have considered cases of *akrasia* without weakness of will. There are some cases that display both. Consider the following three examples.

(a) Ravi has devoted his life to his poetry and still considers it the most important thing in the world. But he has fallen in love with Una, an English schoolgirl. She finds herself pregnant, and, fearing that her father will force her into an abortion, they elope. Ravi vows that no one will harm the child. His commitment is short lived, however. They are found, and the police threaten him with prison. Faced with the choice between standing by Una and the unborn child, or leaving her to a forced abortion and going free to pursue his poetry once again, he shamefacedly chooses the latter.[18]

(b) Christabel, an unmarried Victorian lady, has decided to embark on an affair that she knows will be disastrous. It will ruin her reputation, and quite probably leave her pregnant. Moreover, she considers it morally wrong. So she thinks it not the best option on either moral or prudential grounds. Nevertheless, she has resolved to go ahead with it. At the very last moment, however, she pulls out: not because of a rational reconsideration of the pros and cons, but because she simply loses her nerve.[19]

(c) The President has his finger on the nuclear button and his threat is simple: if the enemy invades, he will retaliate in massive fashion. That will be a catastrophe for everyone, for the President and his people as well as for the enemy. But he reasons that such a threat is needed to deter; and that in order for the threat to be credible it must be genuine: he must really form the conditional intention to retaliate if the enemy invades. Then the unthinkable happens: the enemy invades. Enraged, the President prepares to retaliate. But then, faced with doing what he takes to be morally and prudentially wrong, he hesitates, revises his intention, and desists.[20]

---

[18] Rumer Godden, *The Peacock Spring* (London: Macmillan, 1975).

[19] Thanks to Rae Langton for this example. It is adapted from the story of Christabel LaMotte in A.S. Byatt's novel *Possession* (London: Chatto and Windus, 1990)—Byatt's Christabel, however, does *not* lose her nerve. In thinking about the case, it is important to recall the distinction, made above in discussing the hang-gliding example, between a revision that it is rational to make, and a revision that is made rationally

All three examples have a similar form. At some point, each agent either forms, or finds themselves with, an intention that is in some sense *akratic*: an intention to do other than what they judge best. Ravi intends to stand by Una; Christabel intends to have the affair; the President intends to push the button. None at this point seems to show weakness of will, notwithstanding the gap between what they intend and what they judge best. So here we have *akrasia* without weakness of will; or at least, we would have if the agents had acted on those intentions. Then the agents revise their intentions, bringing them into line with what they judge best. Ravi now intends to abandon Una and pursue his poetry; Christabel now intends to forego the affair; the President now intends to take his finger off the button without pushing it. But in each case, it is precisely the revision that leaves them open to the charge of weakness of will. So here we have weakness of will without *akrasia*.

For Ravi and Christabel, I think it is clear that revising the intention constitutes weakness of will. There are, however, important differences between the two examples. Ravi has a guilty conscience about what he is doing; while he does judge it the best option, all things considered, he does not judge it to be the morally best option. So might we try to save the traditional account by arguing that weakness of will consists in failing to do what one judges to be morally best? Christabel's case shows us that this defense will not work. She does think that the best option, all things considered, is the morally best option, yet in revising her intention she, too, shows weakness of will.[21]

---

[20] I am grateful to David Lewis for this example. I am not sure that he intended it for quite the use to which I have put it.

[21] Advocates of the traditional account have typically insisted that the judgment of what is best, all things considered, is not the same as the judgment of what is morally best; and that the account of weakness of will should be framed in terms of the former notion, not the latter. See Donald Davidson, "How Is Weakness of the Will Possible?" in his *Essays on Actions and Events* (New York: Oxford, 1980), pp. 21-42, here p. 30.

The case of the President is more complicated. It is not clear whether we should ascribe weakness of will to him; but that is because it is unclear whether or not it is reasonable for him to have a tendency to change his mind in this case. Let us return to the point at which the President was still trying to make the deterrence policy work. Then, at least according to the logic of deterrence, it was reasonable for him to form the intention to retaliate if the invasion happened. At that point, we could surely have expressed a conviction that he would stick with this intention by describing him as strong willed; and, equally, we could have expressed a doubt that he would stick with it by describing him as weak willed. That is what the account predicts. But once things have gone wrong, it is not obvious what he should reasonably do, even if we accept the logic of deterrence. On the one hand, he reasonably judged that he should retaliate in case of invasion, and he is now in such a case; on the other, triggering an all-out nuclear war is madness. If we incline to the second of these considerations (as we surely should) then we should not describe him as weak willed if he revises his intention. But the memory of what we said before the invasion should make us uneasy with that judgment. In cases like this, we get ourselves into a very real and very dangerous muddle over what it is reasonable to do; our confusion over applying the notion of weakness of will reflects that muddle.

## IV. RESPONSES TO OBJECTIONS

I have given my reasons in favor of the account that I propose. I shall finish by responding to five possible objections.

*(i) Why be so strict?* Amelie Rorty, in her enlightening paper "Where does the Akratic Break Take Place?" answers her own question by providing a multitude of candidates (*op. cit.*).[22] I have suggested that weakness of will consists in just one kind of failing. But why should I be so strict? The English language is a plastic instrument. It seems overwhelmingly likely that sometimes people have used the expression 'weakness of will' to describe other failings, such as some of the others that Rorty mentions. In particular, is it not very likely that sometimes it has been used to describe the case in which a person fails to form an intention to do what he knows to be best, and as a result does what he knows not to be. In short, is it not very likely that the traditional account captures one of our uses of the expression?

---

[22] Note that Rorty is here giving a discussion of *akrasia*, not of weakness of will; like me, she denies that the two are the same. See note on p. 333.

Perhaps that is right. If so, I should rest content with the claim that many cases of weakness of will are captured by the account proposed here; I should offer it as a supplement to the traditional account, not as a replacement. But I cannot help thinking that the traditional account is not simply inadequate, but straight-out wrong. First, we have seen a number of cases in which people appear to be *akratic* without being weak willed. Second, I doubt that there are any clear cases of weakness of will that can be captured by the traditional account and not by mine.

Consider this example concerning Jim Dixon, hero of Kingsley Amis's novel *Lucky Jim*.[23] Dixon, a junior and untenured lecturer, is staying with Professor Welch, his head of department and arbiter of his future in the university. He awakens after a night of heavy drinking to discover cigarette burns through his bedclothes.

> Had he done all this himself? Or had a wayfarer, a burglar, camped out in his room? Or was he the victim of some Horla fond of tobacco? He thought that on the whole he must have done it himself, and wished he hadn't. Surely this would mean the loss of his job, especially if he failed to go to Mrs. Welch and confess what he'd done, and he knew already that he wouldn't be able to do that (*ibid.*, p. 63).

Would we not say that Dixon displays weakness of will here? Yet he never forms an intention to tell Mrs. Welch. So must we not accept that there can be weakness of will that does not consist of a failure to do what one intends? This seems like good evidence that the traditional account is sometimes correct. But I am not so sure. Why does Dixon not form the intention to tell Mrs. Welch? Presumably, because he knows that he would not be able to go through with it, if he did. So he knows that, if he did form such an intention, he would display weakness of will, where this is understood in terms of the present account. It is because he knows that he is someone with a tendency to weakness of will that he acts as he does. So, on the account given here, his weakness of will explains his action (or rather his inaction). It seems to me that is good enough. Once we have said that we feel no compelling need to insist that Dixon actually exhibits weakness of will here.

If we are to get a really compelling counterexample, we need a clear case of weakness of will in which a person knows that, if he were to form the intention to do what he judges best, he would stick to that intention; but he fails to form the intention. Perhaps I am blinded by partiality to my account, but I am unable to think of one.

---

[23] London: Gollancz, 1953.

*(ii) Can the account accommodate disagreement?* I said that one plausible way of filling out the story of the young Schelling and his blankets is to say that he displayed weakness of will twice over. But there are other things we might say. Knowing Schelling's subsequent career as an economist rather than an explorer, we might say that the real Schelling (the rational realistic Schelling) is the one who wanted to keep the blanket on; hence it was only the fanciful dreamer who displayed weakness of will when trying to get it off. (Contrast Francis Chichester,[24] who, true to his real self, stripped his childhood bed down to just a sheet, seems never to have been tempted to pile on extra blankets, and became a fearless adventurer.) Alternatively, we might say that the weakness of will occurred only in the middle of the night; the real Schelling was the one who made his decisions in the light of day, not tempted by the lure of the warm blanket. We can imagine that debate about this could go on for some time. We can imagine the kinds of evidence that would be relevant. And we can even imagine that, when all the available evidence was in, people might still disagree.

It is easy to see how the traditional account of weakness of will can try to explain this disagreement. The debate would be over what Schelling really valued: his comfortable sleep or his training as an explorer. Can the account offered here explain it as well? I think it can; indeed, I think it can actually do better than the traditional account. The account offered here employs both a descriptive and a normative element. To display weakness of will, an agent must have formed an intention that he revises; that is the descriptive element. And his revision must have been something that, by the standards of a good intender, he should not have done; that is the normative element. Disagreements about whether or when agents manifest weakness of will typically result from disagreements about the second of these two factors. So we might argue that Schelling should have kept to intentions that were formed when he could think clearly (such as in the daytime) rather than when he could not (such as when he was tired at night). Or we might argue that he should have kept to intentions that were realistic (the intention to get a good night's sleep) rather than romantic and unrealistic (the intention to become an Antarctic explorer).

Here, we have the second way in which the vagueness in the account mirrors the vagueness in the concept. I doubt that there will, in general, be agreement on when an agent shows weakness of will; and I suspect that this disagreement is actually better captured by the account on offer here than by the traditional account. On the traditional account, the answers will depend on what the agent thought best, and that is a purely descriptive question. On the account on offer here, the answers will depend on which intentions he should have stuck with as a rational intender. That is a normative question. I suspect that the deep-rootedness of the debate gives us some reason for characterizing it as normative, but I cannot think of a way of resolving the issue with any degree of certainty.

---

[24] *The Lonely Sea and the Sky* (London Hodder and Stoughton, 1964), p 26.

*(iii) Does the account work for policy intentions?* It might be thought that the account cannot accommodate weakness of will with respect to policy intentions. Suppose I intend to get up every morning at six. Then can I not show weakness of will by lounging in bed today till nine, even though I have not revised my policy intention? The right answer is that I have revised the intention. I have not abandoned it altogether, but I have revised it. I have inserted a let-out clause: I intend to get up every day at six, *except today*.

*(iv) Can the account accommodate actions performed without intentions?* I have written as though there is a tight link between intentions and actions: as though we do not get actions without intentions. Perhaps that connection is too tight. Suppose it is. Then it seems that the following is possible. I form an intention to perform some act, and then do something else. But I do not revise my intention; I simply act without an intention. Could that not be a case of weakness of will that does not fit the account offered here?

If there were such cases, I doubt that they would pose too much of a problem; we could extend the account to embrace certain failures to act on intentions as well as revisions of intentions. (Not all such failures should be included: I do not typically display weakness of will if I *forget* my intention.) But I am not convinced that there are any cases of this kind. Let us consider one of Bratman's examples of action without intention. Suppose someone throws me a ball and, without thinking, I just reach out and catch it. That is perhaps an intentional act; but it is not clear that I formed an intention to catch it. I acted quite without reflection.[25] But now suppose I form the intention not to catch the ball; you throw it, and again I just reach out and catch it. Have I shown weakness of will? I rather doubt that I have. I suspect that in so far as the act really is one that is done without an intention, it is one that falls outside my immediate control, and so outside of the area in which weakness of will can be shown. I might resolve not to faint when the incision is made, not to duck when the jets fly low overhead, not to smile when my friend inadvertently makes a clown of himself. If I cannot help myself do these things, however, my failure to stick by my resolution does not manifest weakness of will. Certainly, these are failures of self-control. But this is just another reason for not equating such failure with weakness of will.

---

[25] Bratman, *Intention, Plans, and Practical Reason*, p. 126. He also discusses actions that do not result from an intention but from what he calls an "endeavoring" (chapter 9). If we follow him in thinking that the distinction is necessary, shall we need to augment the present account by substituting 'intention or endeavoring' in place of each occurrence of 'intention'? I am not sure, but I think not, since, on Bratman's account, each endeavoring to do something which is not an intention to do it will nonetheless be a intention to do some other thing (p. 119), so each over hasty revision of an endeavoring will also be an over hasty revision of an intention.

*(v) Can the account distinguish weakness of will from compulsion?* It is a commonplace in the philosophical literature that weakness of will should be distinguished from compulsion. On the traditional account, the distinction is made as follows: one displays weakness of will when one *freely* acts against one's own best judgment; one is compelled when one's action against one's own best judgment is *not free*.[26] But it might seem that the account offered here can make no such distinction. Consider a compulsive person—a kleptomaniac, for instance—who intends never to steal again, but who revises that intention and does. Will his revision not be one that he rationally should not have made? Yet if so, he will count as weak willed.[27]

I have two things to say about this. First, it is not obvious to me that it would be a disaster if the account did classify compulsive acts as weak willed; for it is not obvious to me that they are not. Certainly, we would not normally *say* that a compulsive person was weak willed, but that could be because it would be inappropriate rather than false to say so—in the way that it would be inappropriate to say that sadistic torturers are unkind. Are we really averse to saying that compulsives are *pathologically* weak willed?

Second, it is not obvious whether the account offered here would classify compulsives as weak willed. Certainly, one can imagine cases for whom the compulsive action is so automatic that there is no need of an intention, and hence no need to revise any contrary intention: for the kleptomaniac who simply finds himself placing objects in his pocket, intentions to steal or not to steal are quite beside the point. So such cases will not be classed as weak willed. But what of the kleptomaniac who consciously but compulsively does revise his intention not to steal; that is, who is simply unable to refrain from making the revision? Will he count as someone who has an unreasonable tendency to revise, and hence as someone who is weak willed? This depends, of course, on how we determine what is reasonable. We could use internal criteria, criteria that always restrict what it is reasonable for an agent to do to that which it is in the agent's power to do.[28] Such criteria would not classify the kleptomaniac's revision as unreasonable, and hence would not classify him as weak willed. Alternatively, we could use external criteria, criteria that do not restrict what it is reasonable for an agent to do to that which it is in the agent's power to do. Such criteria probably would class the kleptomaniac's revision as unreasonable, and hence probably would classify him as weak willed. I have not come down in favor of one or the other of these criteria of what is reasonable; indeed, I suspect that while I am trying to give an account of our ordinary notion of weakness of will, I should remain uncommitted. For I suspect that this is another place where our ordinary concept is vague.

---

[26] See Watson; Kennett and Michael Smith, "Philosophy and Commonsense: The Case of Weakness of Will," in Michaelis Michael and John O'Leary-Hawthorn, eds., *Philosophy in Mind* (Dordrecht: Kluwer, 1994), pp. 141-57. Watson argues that the relevant notion of a free action is one which the agent could have performed had they developed in themselves the kinds of skills and capacities that we expect them to develop.

[27] Thanks to Smith and Linda Barclay for making me think about this issue.

So the present account of weakness of will has difficulties with compulsion only if two further claims are met: if it is right to use external criteria of what is reasonable, *and* if the compulsive should certainly not count as weak willed. I have questioned both of these claims; but even if they were true, a simple amendment to the account would fix things up. Rather than say that weakness of will consists in overreadily revising an intention, say that it consists in overreadily revising an intention *when it is in the agent's power to desist from that revision.*

RICHARD HOLTON

University of Sheffield

---

[28] In the agent's power in some broad sense; we could follow Watson in thinking that what is important is what would have been in the agent's power if the agent had developed the capacities that we normally expect them to develop.

## BOOK REVIEWS

*Open Minded: Working Out the Logic of the Soul.* JONATHAN LEAR. Cambridge: Harvard University Press, 1998. 342 p. Cloth $35.00.

A mind that is not open is closed; it is a closed *mind*. So the difference between them cannot be that the one does, while the other does not, entertain ideas. Nor can it lie in the heat of the entertaining, since open-mindedness is consistent with passionate conviction; it depends on how conviction is achieved, on whether it is sustained by self-deception. Perhaps, then, only one who knows his own mind can have a mind that is open.

In the preface to his provocative, often brilliant book, Jonathan Lear writes: "We lose sight of Plato's project, laid out so beautifully in the *Republic*, of giving a *non*formal yet not nonempirical account of what it is to be human. Plato, one might say, is working out [in the *Republic*] *the very idea* of what it is to be minded as we are. And he does this in the light of Socrates' exemplification—a life spent showing us—that one of the most important truths about us is that we have the capacity to be *open-minded*: the capacity to live nondefensively with the question of how to live" (8). It is a story that is best continued later by Sigmund Freud, for it is he who teaches us about the mind's strategies of defense, about the varieties of psychic pain and psychic trauma, and about how our defenses against them may sometimes be more damaging to the mind than the enemy it tries to keep out.

Plato and Freud are the two intellectual heroes of this book, the twelve essays of which together yield a vision that is Freudian in the best sense. The dark side of the vision is that the mind is not so rational—in that sense not so open—as we like to think; that the division between the known and the unknown comes not where mind edges world, but within the mind itself; and that defensively to deny reason's limits invites both personal and social disaster. The lighter side is that resistances to reason are not always final, and that a greater degree of psychic wholeness, which the mind naturally wants, can be gained precisely by working with and through resistance.

The lead essay titled "On Killing Freud (Again)," which first appeared in *The New Republic* as a response to the Library of Congress's decision to cancel the Freud exhibit, sets the tone. Like most psychoanalysts, Lear interprets the Oedipal complex as having to do with childhood phantasies of omnipotence. The complex is "resolved" in relinquishing these phantasies, so acknowledging the limits of both

0022-362X/99/9605/263–69

one's power and one's knowledge. This is the sense in which, on Lear's reading, the drama of Sophocles' play is *Oedipal.* Oedipus manically insists that he is a self-made man, sufficient unto himself, that prophecy is blind, reason is all. Lear calls these pretensions *knowingness,* a defensive attitude motivated by anxieties about one's limits. Whether we call the unknown "Fate" or "the unconscious," tragic consciousness in any culture consists in the recognition that human beings live their lives, as Lear puts it, "in the intersection of two realms of meaning" (53). He grants that the Freud-bashers have many legitimate complaints, but claims one can dismiss Freud out of hand only by denying that we are often moved by forces within us which we do not understand. The Freud-bashers exhibit a contemporary form of knowingness.

The companion essay that follows, "Knowingness and Abandonment: An Oedipus for Our Time," develops these ideas into a critique of postmodernist thinking. The Enlightenment "constituted itself with a manic, oedipal defense: even though the gods have left us, human reason can take their place" (54). But since its function is to defend the mind against something awful that it half recognizes to be the case, such a defense is inherently unstable. Postmodern consciousness, Lear suggests, originates in the collapse of this defense. The idea is not that the postmodernist has worked through his "Oedipal" problems about knowing, rather that one of the criteria for a "successful" analysis, as it were, is the ability to recognize and face the problems as they continually reappear. The psychoanalytic aim is not to get rid of all defenses, but only those which are inflexible, outdated, and, on the whole, too costly.

In "Restlessness, Phantasy, and the Concept of Mind," Lear argues that philosophical tradition in the West has often held a view of the mind skewed toward reason in ways Freud tried to correct. Lear's strategy is first to show that a familiar contemporary account of irrationality, which I shall call *the rationalizing account,* is unsatisfactory; then to show, in giving a better account of it, that what we call irrationality is not a knot at the mind's brim, but a lead into its very center.

On the rationalizing account, as Lear presents it, irrationality requires us to posit a mind subdivided into structures of propositional attitudes, each of which is rationally coherent within itself. "On this schema, the unconscious mind would be conceived of as its own locus of rationality (or quasi-rationality)..." (82). The account holds that in this way we can allow for that degree of holism and rationality which the concept of the mental presumes, while making room for phenomena like self-deception and *akrasia.* The move that makes for division, the account continues, is a mental cause, namely, the wish to ignore the judgment of reason in a particular case, that is not a reason (a good reason) for what it causes. Lear clearly has Donald Davidson in mind in all this, though he mentions him by name only in the notes.

Lear thinks the rationalizing account unsatisfactory because it leaves unexplained the fact of irrationality itself, that is, why in any particular case the better reason was not operative. It teaches us nothing about how the irrationality comes into being. Its fundamental error is the failure to see that restlessness, as Lear puts it, is fundamental to the mind, an idea he thinks was one of Freud's distinctive achievements. In the Freudian view, restlessness expresses itself in *mental tropisms,* like *projective* and *introjective* phantasies, through which the mind attempts to rid itself of anxiety; *the drives* (of which sexuality is the paradigm); and *primary-process mental activity.* (Lear uses the ph spelling of 'phantasy' to differentiate it from the ordinary concept of fantasy as imagining or daydreaming.) All these concepts are part of the basic Freudian repertoire, though Lear's interpretation of them is influenced both by Melanie Klein and Hans Loewald (about whom I say more in a moment).

Here is my own rough explication of projection. Let us say that a child is furious with someone on whom he is vitally dependent. Since he may then be at some level aware of his own anger as endangering this relationship, he becomes anxious. The mind may then handle this anxiety by constructing an affective picture or an affective (nonpropositional) thought of the anger as in another person, or as part of another person. The anger is cast out of one's own mind, so to speak, and projected onto, or (in one's imagination) into, the other.

Lear does not want us to think of this manoeuver as fully intentional, but rather as instancing a mental tropism, which he defines as "a subintentional, nonrational mental activity which has a strategic point: namely, relieving anxiety" (118). Phantasy disrupts rational thought processes in a way which is purposive but which cannot be accommodated to the model of akrasia or self-deception. (This idea was first articulated by Richard Wollheim in *The Thread of Life,* and picked up both by James Hopkins and myself.[1] Phantasies must be seen, Lear argues, not only as contents *in* the mind, but also as acts *upon* the mind—for example, dividing a mind into parts, or keeping it from achieving the sort of integration that rationality presumes. The idea of division is common to both Lear and the rationalizing account. But Lear envisions the act of dividing and the divided portions themselves differently. The concept of phantasy helps articulate these differences.

---

[1] Wollheim, *The Thread of Life* (New York: Cambridge, 1984); Hopkins, Introduction to *Philosophical Essays on Freud* (New York: Cambridge, 1982); and my *The Psychoanalytic Mind: From Freud to Philosophy* (Cambridge: Harvard, 1993).

In commenting on this, I now want to go back to the rationalizing account. I quoted Lear as saying that it treats irrationality "as a *by-product* of the mind's being a composite of two quasi-minds." Lear apparently sees division on the rationalizing account as just a fact about the mind, from which irrationality somehow ensues. This mistakes Davidson's argument in an important way: irrationality on his view is not a by-product of division, but the very phenomenon that requires us, as philosophical analysts, to posit division. Davidson grants, furthermore, that there are many phenomena that are neither rational nor irrational, but arational *and* mental (perhaps phantasy belongs here). Finally, he is explicit in saying that his account is mute about the kind of psychological detail Lear is after. It is mute on the descriptive, psychological question of what goes on in the mind when it is in the process of dividing, and how it looks before and after the division. Lear correctly sees that, by Davidson's own admission, the rationalizing account is highly abstract.

Davidson also explicitly rejects the idea Lear imputes to him that the divisions in irrationality should be described as conscious versus unconscious. Neither division is entirely rational in the way Lear suggests. Instead, in Davidsonian division there is room *in both parts of the divided mind* for contents that are not the sort of thing that are reasons in any case, and for connections between ideas that are neither rational nor irrational. Lear comments that with "so much rationality seemingly built into the very idea of mind, it's a wonder we can ever take an irrational breath" (83). Taking a breath is not something we usually do either for or in spite of reason. It is usually neither irrational nor rational. In urging that philosophy of mind needs more distinctions than philosophers have given it, Lear overlooks some it has.

But let us suppose that Lear's description of what is missing in the rationalizing account is more or less right. Then we might ask the following questions. Is the rationalizing account compatible with Lear's description? Even if it is, does this description nevertheless suggest a picture of the mind that is substantially different from that of the rationalizing account? How can Lear's picture accommodate the larger considerations about meaning, language, and communication which he does not address? And is it the case, as Lear claims but does not really argue, that the sources of irrationality feed also those forms of creativity without which there would be no mind at all of the sort we value?

These are hard questions I do not know how to answer. This much seems clear: the difference between Lear's position and that of the rationalizing account is at least a matter of temperament; and, like all such differences, it leads to differences in emphasis, even in perception. Like Lear, I believe that the psychoanalytic data force one to a concept of psychic intelligibility that escapes the concept of rationality as we often think of it. Whether his views are right in the detail, Lear illumines this area of ambiguity between the rational, with its variant in irrationality, and the arational.

Only two of the essays in the book have to do with subjects that are strictly psychoanalytic. The first is an interpretation of the psychoanalytic concept of transference. Readers of Freud usually understand it to refer to a specific kind of thought process: the carrying over onto a present interpersonal relationship of those perceptions, feelings, and behavior which had their origin in a past relationship and which tend to stamp the present with a mold that is rigidly familiar and deeply conflicted. Lear discusses transference in this sense. But often he speaks of it as if it were more or less synonymous with thought in general. He says that "transference...is just the psychological characteristic activity of creating a meaningful world in which to live" (60); and that as it emerges in analysis, transference "is simply the idiosyncratic, unconscious side of the psyche's fundamental activity: to inform the world with meaning" (70-71). The key to Lear's interpretation is in the work of Loewald, a remarkable American psychoanalyst to whom Lear acknowledges an important debt. For philosophers interested in the turns taken in recent psychoanalytic thinking by the concepts of instinct, Eros, Thanatos, and the repetition compulsion, Lear's other "psychoanalytic" essay, "The Introduction of Eros: Reflections on the Work of Hans Loewald," should be invaluable.

My favorite essay is the one titled "Eros and Unknowing: The Psychoanalytic Significance of Plato's *Symposium*." On Lear's reading, the dialogue begins as a comedy, and ends as a tragedy, with Alcibiades simply unable to transcend his particularity, his bodily nature, as Socrates, through Diotima, wants him to do. Alcibiades, longing and forlorn, the very embodiment of Aristophanes' tragicomic fable about our cloven ancestors, is left behind in the gap between here and there. Lear writes: "By now it should be clear that the psychoanalytic approach to the human-erotic differs fundamentally from the Socratic. For one thing, psychoanalytic 'ascent' is toward individuation rather than away from it" (165). And the body, as it figures, for example, in phantasy, enhances individuation. Eros was a divine source of ascent for Socrates, but only a distraction to Alcibiades. "From a psychoanalytic perspective, by contrast, this erotic distraction needs to be worked with and worked through [to be seen as a 'transference' insistence that another conform to set roles] if human life is to be afforded a comic restoration" (165-66). This is a possibility which Lear believes Plato, but not Socrates, saw.

A comparison between psychotherapy and philosophy easily suggests itself to philosophers, but it has never been developed as deftly as it is in these essays. In "Transcendental Anthropology," Lear works out a subtle and complex argument to which I cannot do justice here, an argument that culminates in an analogy between the Wittgensteinian philosopher, examining "our" form of life, and the psychoanalytic patient's process of self-discovery. The analogy takes off from a discussion of the difference in Immanuel Kant between the pure apperception of consciousness and the 'I think': the first must be able to accompany all representations, it informs us about the 'I' that is thinking, but, unlike the 'I think', it cannot itself be represented or observed. Similarly, the Wittgensteinian philosopher learns that "'observing' the context in which rule-following activity is embedded will not yield the full meaning of that activity..." (273). Such a philosopher is trying to gain "nonobservational" insight into something—he calls it a form of life—that is too intimate a part of our observations to be itself observed. Similarly again, although the analysand learns to become a better observer of himself, the point of therapy "is not the observation of an accurately represented person, but nonobservational insight into the person creating the representation" (272).

This raises an interesting question about just how we should understand the psychoanalytic concept of the unconscious. Lear does not take this up, but looks instead in this essay toward philosophy. He suggests that the familiar complaint that psychoanalysis is not an empirical discipline may be a reason for hope, not disdain; "for perhaps reflection on the therapeutic model will shed light on how to proceed philosophically in a way which is neither a methodology of the sciences nor a purely transcendental investigation" (272).

I have one large complaint about this book, and that is its promiscuous use of the word 'meaning'. It is an idea at the heart of the essays on transference and irrationality. Yet passages like the following, about dreams and dream interpretation, are about as close to an analysis of meaning itself as Lear ever comes: "How could mind be making a meaning it doesn't understand? To be making a meaning, it must be making certain associations among ideas, engaging in symbolization, however elementary" (85). And what is this process of "making a meaning," a phrase Lear uses more than once? Is it, in particular, as Lear suggests, something one does all by oneself? Such promiscuity is particularly noticeable in a book that invokes Ludwig Wittgenstein at the beginning (the preface speaks briefly of Wittgenstein's having dispelled "the myth of meaning") and devotes to him its two concluding essays.

But perhaps the best way to read *Open Minded* is by putting yourself in its hands. Then you will have the intense pleasure of traversing a terrain you thought you knew by a new road, from which, rounding a bend, you have again and again a stunning glimpse of a familiar yet usually hidden landscape. In the Sophoclean universe, Lear remarks, there is either a rigid reliance on human reason, *or* submission to a divine realm. "In neither position is there room for philosophy, that peculiarly thoughtful response to awe" (51). *Open Minded* is itself such a philosophical response.

MARCIA CAVELL

Berkeley, California

## NEW BOOKS

HENRY, PHILIP R.: *Belief Reports and the Structure of Believing.* Stanford: CSLI, 1998. vii, 103 p. Cloth $49.95, paper $18.95.

HOLLAND, NANCY J.: *The Madwoman's Reason: The Concept of the Appropriate in Ethical Thought.* University Park: Penn State UP, 1998. xxxi, 119 p. Cloth $30.00, paper $14.95.

HOLLIS, MARTIN: *Trust within Reason.* New York: Cambridge, 1998. viii, 170 p. Cloth $54.95, paper $16.95.

HOLT, ROBIN: *Wittgenstein, Politics, and Human Rights.* New York: Routledge, 1997. 172 p. Cloth $60.00.

HÖSLE, VITTORIO: *Objective Idealism, Ethics, and Politics.* Notre Dame: University Press, 1998. viii, 227 p. Cloth $29.95, paper $18.00.

HOUGH, SHERIDAN: *Nietzsche's Noontide Friend: The Self as Metaphoric Double.* University Park: Penn State UP, 1997. xxv, 158 p. Cloth $35.00.

HOWELLS, CHRISTINA: *Derrida: Deconstruction from Phenomenology to Ethics.* Malden, MA: Polity/Blackwell, 1999. xi, 175 p. Cloth $54.95, paper $21.95.

HOY, TERRY: *The Political Philosophy of John Dewey.* Westport, CT: Greenwood, 1998. 141 p. Cloth $49.95.

HUNT, LESTER H.: *Character and Culture.* Lanham, MD: Rowman and Littlefield, 1997. xii, 302 p. Cloth $67.50, paper $24.95.

HURLEY, S.L.: *Consciousness in Action.* Cambridge: Harvard, 1998. xii, 506 p. Cloth $55.00.

IMBO, SAMUEL OLUOCH: *An Introduction to African Philosophy.* Lanham, MD: Rowman and Littlefield, 1998. xvii, 157 p. Cloth $53.00, paper $19.95.

JACKSON, FRANK: *From Metaphysics to Ethics: A Defence of Conceptual Analysis.* New York: Oxford, 1998. ix, 174 p. Cloth $35.00.

JAMES, SUSAN: *Passion and Action: The Emotions in Seventeenth-Century Philosophy.* New York: Oxford, 1997. x, 318 p. Cloth $35.00.

JANTZEN, GRACE M.: *Becoming Divine: Towards a Feminist Philosophy of Religion.* Bloomington: Indiana UP, 1999. viii, 296 p. Cloth $49.95, paper $24.95.

JENKINS, JOHN I.: *Knowledge and Faith in Thomas Aquinas.* New York: Cambridge, 1997. xiii, 267 p. Cloth $59.95.

JOHANSEN, T.K.: *Aristotle on the Sense-Organs.* New York: Cambridge, 1998. xvi, 304 p. Cloth $59.95.

JOHNSON, BARBARA: *The Feminist Difference: Literature, Psychoanalysis, Race, and Gender.* Cambridge, Harvard, 1998. viii, 215 p. Cloth $22.00.

JONES, JUDITH A.: *Intensity: An Essay on Whiteheadian Ontology.* Nashville: Vanderbilt, 1998. xiv, 247 p. Cloth $39.95.

KATZ, JERROLD L.: *Realistic Rationalism.* Cambridge: MIT, 1998. xxxiv, 226 p. Cloth $35.00.

KEARNEY, RICHARD: *Poetics of Imagining: Modern to Postmodern.* New York: Fordham, 1998. ix, 260 p. Cloth $35.00, paper $19.95.

KEKES, JOHN: *A Case for Conservatism.* Ithaca: Cornell, 1998. x, 239 p. Cloth $28.50.

THE JOURNAL OF PHILOSOPHY
VOLUME XCVI, NO. 6, JUNE 1999

UNIVERSITY
ALIPUR
CAMPUS
8-1-7

# INTERNALISM EXPOSED*

In recent decades, epistemology has witnessed the development and growth of externalist theories of knowledge and justification.[1] Critics of externalism have focused a bright spotlight on this approach and judged it unsuitable for realizing the true and original goals of epistemology. Their own favored approach, internalism, is defended as a preferable approach to the traditional concept of epistemic justification.[2] I shall turn the spotlight toward internalism and its most prominent rationale, revealing fundamental problems at the core of internalism and challenging the viability of its most popular rationale. Although particular internalist theories such as (internalist) foundationalism and coherentism will occasionally be discussed, those specific theories are not my primary concern.

---

* An earlier version of this paper was presented in Pittsburgh, at the Central Division meeting of the American Philosophical Association, April 25, 1997. My commentator on that occasion was Matthias Steup, and I am much indebted to him for valuable correspondence on this topic. I am also grateful to Tim Bayne and Holly Smith for very useful suggestions.

[1] Prominent statements of externalism include D. M. Armstrong, *Belief, Truth and Knowledge* (New York: Cambridge, 1973); Fred Dretske, *Knowledge and the Flow of Information* (Cambridge: MIT, 1981), Robert Nozick, *Philosophical Explanations* (Cambridge: Harvard, 1981); my *Epistemology and Cognition* (Cambridge: Harvard, 1986); and Alvin Plantinga, *Warrant and Proper Function* (New York: Oxford, 1993).

[2] Major statements of internalism include Roderick Chisholm, *Theory of Knowledge* (Englewood Cliffs, NJ: Prentice-Hall, 1966, 1st edition; 1977, 2nd edition; 1989, 3rd edition); Laurence BonJour, *The Structure of Empirical Knowledge* (Cambridge: Harvard, 1985); John Pollock, *Contemporary Theories of Knowledge* (Totowa, NJ: Rowman and Littlefield, 1986); Richard Foley, *The Theory of Epistemic Rationality* (Cambridge: Harvard, 1987); and Keith Lehrer, *Theory of Knowledge* (Boulder: Westview, 1990). In addition to relatively pure versions of externalism and internalism, there are also mixtures of the two approaches, as found in William Alston, *Epistemic Justification* (Ithaca: Cornell, 1989); Ernest Sosa, *Knowledge in Perspective* (New York: Cambridge, 1991); and Robert Audi, *The Structure of Justification* (New York: Cambridge, 1993)

0022-362X/99/9606/271-93

The principal concern is rather the general architecture of internalism, and the attempt to justify this architecture by appeal to a certain conception of what justification consists in.

## I. DEONTOLOGY, ACCESS, AND INTERNALISM

I begin with a certain rationale for internalism that has widespread support. It can be reconstructed in three steps:

(1) The *guidance-deontological* (GD) *conception of justification* is posited.
(2) A certain constraint on the determiners of justification is derived from the GD conception, that is, the constraint that all justification determiners must be *accessible to*, or *knowable by*, the epistemic agent.
(3) The accessibility or knowability constraint is taken to imply that only internal conditions qualify as legitimate determiners of justification. So justification must be a purely internal affair.[3]

What motivates or underlies this rationale for internalism? Historically, one central aim of epistemology is to guide or direct our intellectual conduct, an aim expressed in René Descartes's title, "Rules for the Direction of the Mind."[4] Among contemporary writers, John Pollock expresses the idea this way:

> I have taken the fundamental problem of epistemology to be that of deciding what to believe. Epistemic justification, as I use the term, is concerned with this problem. Considerations of epistemic justification guide us in determining what to believe. We might call this the "belief-guiding" or "reason-guiding" sense of 'justification' (*op. cit.*, p. 10).

The guidance conception of justification is commonly paired with the deontological conception of justification. John Locke[5] wrote of a person's "duty as a rational creature" (*ibid.*, p. 413), and the theme of epistemic duty or responsibility has been echoed by many contemporary epistemologists, including Laurence BonJour (*op. cit.*), Roderick Chisholm (*op. cit.*), Carl Ginet, Paul Moser, Matthias Steup, Richard Feldman, and Hilary Kornblith.[6] Chisholm defines cousins of the concept of justification in terms of the relation 'more reasonable than', and he re-expresses the relation 'p is more reasonable than q for S at t' by saying: "S is so situated at t that his intellectual *requirement*, his *responsibility* as an intellectual being, is better fulfilled by p than by q."[7] Similarly, Feldman says that one's epistemic duty is to "believe what is supported or justified by one's evidence and to avoid believing what is not supported by one's evidence" (*op. cit.*, p. 254).

---

[3] Plantinga also traces internalism to the deontological conception: "If we go back to the source of the internalist tradition,...we can see that internalism arises out of deontology; a deontological conception of warrant...leads directly to internalism" (*op. cit.*, pp. 24-25). Alston proposes a slightly different rationale for internalism, although his rationale also proceeds via the knowability constraint (*op. cit.*, p. 236). He suggests that the concept of justification derives from the interpersonal practice of criticizing one another's beliefs and asking for their credentials. A person can appropriately respond to other people's demands for credentials only if he knows what those credentials are. So it is quite understandable, says Alston, that justifiers must meet the requirement of being accessible to the agent. Clearly, this is one way to derive the accessibility constraint without appeal to the deontological conception. But Alston is the only one I know of who advances this ground for the accessibility constraint. In any case, most of the problems I shall identify pertain to the accessibility constraint itself, which Alston's rationale shares with the deontological rationale.

[4] *Philosophical Works of Descartes, Volume I*, Elizabeth Haldane and G. R. T. Ross, trans. (New York: Dover, 1955)

[5] *An Essay Concerning Human Understanding, Volume II*, A. C. Fraser, ed. (New York: Dover, 1955).

The guidance and deontological conceptions of justification are intimately related, because the deontological conception, at least when paired with the guidance conception, considers it a person's epistemic duty to guide his doxastic attitudes by his evidence, or by whatever factors determine the justificational status of a proposition at a given time. Epistemic deontologists commonly maintain that being justified in believing a proposition *p* consists in being (intellectually) required or permitted to believe *p*; and being unjustified in believing *p* consists in not being permitted, or being forbidden, to believe *p*. When a person is unjustified in believing a proposition, it is his duty not to believe it.

It is possible to separate the deontological conception from the guidance idea. In ethical theory, a distinction has been drawn between accounts of moral duty that aim to specify what makes actions right and accounts of moral duty that aim to provide practical decision procedures for what to do.[8] If an account simply aims at the first desideratum, it need not aspire to be usable as a decision guide. Similarly, accounts of epistemic duty need not necessarily be intended as decision guides. When the deontological conception is used as a rationale for epistemic internalism of the sort I am sketching, however, it does incorporate the guidance conception. Only if the guidance conception is incorporated can the argument proceed along the intended lines to the accessibility constraint, and from there to internalism. This is why I shall henceforth speak of the GD conception of justification.

---

[6] Ginet, *Knowledge, Perception, and Memory* (Dordrecht Reidel, 1975); Moser, *Empirical Justification* (Dordrecht: Reidel, 1985); Steup, "The Deontic Conception of Epistemic Justification," *Philosophical Studies*, LIII (1988): 65-84; Feldman, "Epistemic Obligations," in J. Tomberlin, ed., *Philosophical Perspectives*, Volume II (Atascadero, CA: Ridgeview, 1988), pp. 235-56; and Kornblith, "Justified Belief and Epistemically Responsible Action," *Philosophical Review*, XCII (1983)· 33–48

[7] *Theory of Knowledge*, 2nd edition, p. 14 (emphasis added).

[8] For example, R. Eugene Bales distinguishes between two possible aims of act-utilitarianism: as a specifier of a right-making characteristic or as a decision-making procedure. See "Act-utilitarianism: Account of Right-making Characteristics or Decision-making Procedure," *American Philosophical Quarterly*, VIII (1971). 257-65. He defends utilitarianism against certain critics by saying that it does not *have* to perform the latter function.

I turn now to the second step of the argument for internalism. Following William Alston,[9] I shall use the term *justifiers* for facts or states of affairs that determine the justificational status of a belief, or the epistemic status a proposition has for an epistemic agent. In other words, justifiers determine whether or not a proposition is justified for an epistemic agent at a given time. It seems to follow naturally from the GD conception of justification that a certain constraint must be placed on the sorts of facts or states of affairs that qualify as justifiers. If a person is going to avoid violating his epistemic duty, he must know, or be able to find out, what his duty requires. By *know*, in this context, I mean only: have an *accurate*, or *true*, belief. I do not mean: have a *justified* true belief (or whatever else is entailed by the richer concept of knowledge). Admittedly, it might be possible to avoid violating one's duties by chance, without knowing (having true beliefs about) what one's duties are. As a practical matter, however, it is not feasible to conform to duty on a regular and consistent basis without knowing what items of conduct constitute those duties. Thus, if you are going to choose your beliefs and abstentions from belief in accordance with your justificational requirements, the facts that make you justified or unjustified in believing a certain proposition at a given time must be facts that you are capable of knowing, at that time, to hold or not to hold. There is an intimate connection, then, between the GD conception of justification and the requirement that justifiers must be accessible to, or knowable by, the agent at the time of belief. If you cannot accurately ascertain your epistemic duty at a given time, how can you be expected to execute that duty, and how can you reasonably be held responsible for executing that duty?[10]

The *knowability constraint on justifiers* which flows from the GD conception may be formulated as follows:

> KJ: The only facts that qualify as justifiers of an agent's believing $p$ at time $t$ are facts that the agent can readily know, at $t$, to obtain or not to obtain.

---

[9] "Internalism and Externalism in Epistemology," reprinted in his *op. cit.*, pp. 185-226, here p. 189.

[10] Some internalists explicitly reject externalism on the grounds that it cannot be used as a decision guide. For example, Pollock says: "[I]t is in principle impossible for us to actually employ externalist norms. I take this to be a conclusive refutation of belief externalism" (*op. cit.*, p. 134). He would not subscribe to the full argument for internalism I am discussing, however, because it is committed to the "intellectualist model" of epistemology, which he disparages.

How can an agent readily know whether candidate justifiers obtain or do not obtain? Presumably, the agent must have a way of determining, for any candidate class of justifiers, whether or not they obtain. Such a way of knowing must be reliable, that is, it must generate beliefs about the presence or absence of justifiers that are usually (invariably?) correct. Otherwise, the agent will often be mistaken about what his epistemic duty requires. The way of knowing must also be "powerful," in the sense that when justifiers obtain it is likely (certain?) that the agent will believe that they obtain; at least he will believe this if he reflects on the matter or otherwise inquires into it.[11] As we shall soon see, internalists typically impose additional restrictions on how justifiers may be known. But the minimal, generic version of KJ simply requires justifiers to be the sorts of facts that agents have *some* way of knowing. In other words, justification-conferring facts must be the sorts of facts whose presence or absence is "accessible" to agents.[12]

Given the KJ constraint on justifiers, it becomes fairly obvious why internalism about justification is so attractive. Whereas external facts are facts that a cognitive agent might not be in a position to know about, internal facts are presumably the sorts of conditions that a cognitive agent can readily determine. So internal facts seem to be the right sorts of candidates for justifiers. This consideration leads to the third step of our rationale for internalism. Only internal facts qualify as justifiers because they are the only ones that satisfy the KJ constraint; at least so internalists suppose.

One possible way to criticize this rationale for internalism is to challenge the GD conception directly. This could be done, for example, by arguing that the GD conception of justification presupposes the dubious thesis of doxastic voluntarism, the thesis that doxastic attitudes can be "guided" by deliberate choices or acts of will. This criticism is developed by Alston,[13] and I have sympathy with many of his points. But the voluntarism argument against the GD conception is disputed by Feldman (*op. cit.*) and John

---

[11] For the distinction between reliability and power (phrased slightly differently), see my *op. cit.*, chapter 6

[12] Jack Lyons points out that to comply with one's epistemic duty it suffices to know *that one has* (undefeated) justifiers for proposition *p*; one does not have to know *which* justifiers these are So the argument is not entitled to conclude that knowledge of particular justifiers is required by epistemic duty. Practically speaking, however, it is difficult to see how a cognitive agent could know that relevant justifiers exist without knowing which particular ones exist. So I shall pass over this objection to the internalist line of argument.

[13] "The Deontological Conception of Justification," reprinted in his *op. cit.*, pp. 115-52.

Heil,[14] among others. Feldman, for example, argues that epistemic deontologism is not wedded to the assumption of doxastic voluntarism. Many obligations remain in force, he points out, even when an agent lacks the ability to discharge them. A person is still legally obligated to repay a debt even when his financial situation makes him unable to repay it. Perhaps epistemic obligations have analogous properties.[15] Since the complex topic of doxastic voluntarism would require article-length treatment in its own right, I set this issue aside and confine my attention to other issues. Although I do not accept the GD conception of justification, I take it as given for purposes of the present discussion and explore where it leads. In any case, what is ultimately crucial for internalism is the accessibility requirement that the GD conception hopes to rationalize. Even if the GD conception fails to provide a good rationale, internalism would be viable if some other rationale could be provided for a suitable accessibility requirement.

## II. DIRECT KNOWABILITY AND STRONG INTERNALISM

The initial KJ constraint was formulated in terms of knowability plain and simple, but proponents of internalism often add the further qualification that determinants of justification must be *directly* knowable by the cognitive agent. Ginet, for example, writes as follows:

> Every one of every set of facts about S's position that minimally suffices to make S, at a given time, justified in being confident that p must be *directly recognizable* to S at that time (*op. cit.*, p. 34).

Similarly, Chisholm writes:

> [T]he concept of epistemic justification is...internal and immediate in that one can *find out directly*, by reflection, what one is justified in believing at any time.[16]

Thus, Ginet and Chisholm do not endorse just the minimal KJ constraint as earlier formulated, but a more restrictive version, which might be written as follows:

---

[14] "Doxastic Agency," *Philosophical Studies*, XL (1983): 355-64

[15] Feldman's response, however, undercuts the step from the GD conception of justification to the knowability constraint. If epistemic duty does not require that the agent be *able* to discharge this duty, there is no longer a rationale for the knowability constraint. A different line of response to the voluntarism worry is taken by Lehrer, who suggests that epistemological analysis should focus not on belief but on *acceptance*, where acceptance is some sort of action that is subject to the will—"A Self-Profile," in R. Bogdan, ed., *Keith Lehrer* (Dordrecht: Reidel, 1981), pp. 3-104.

[16] *Theory of Knowledge*, 3rd edition, p. 7, emphasis added and original emphasis deleted

$KJ_{dir}$: The only facts that qualify as justifiers of an agent's believing *p* at time *t* are facts that the agent can readily know *directly*, at *t*, to obtain or not to obtain

An initial problem arising from $KJ_{dir}$ is this: What warrants the imposition of $KJ_{dir}$ as opposed to the looser constraint, KJ? KJ was derived from the GD conception on the grounds that one cannot reasonably be expected to comply with epistemic duties unless one knows what those duties are. How does such an argument warrant the further conclusion that *direct* knowledge of justification must be available? Even indirect knowledge (whatever that is) would enable an agent to comply with his epistemic duties. So the second step of the argument for internalism cannot properly be revised to feature $KJ_{dir}$ in place of KJ. Proponents of $KJ_{dir}$ might reply that direct forms of knowledge are more powerful than indirect knowledge, but this reply is unconvincing. The power requirement was already built into the original version of KJ, and it is unclear how directness adds anything of significance on that score. Whether $KJ_{dir}$ can be derived from GD is a serious problem, because the argument for internalism rests on something like the directness qualification. I shall say more about this later; for now I set this point aside in order to explore where $KJ_{dir}$ leads.

What modes of knowledge count as direct? At least one form of direct knowledge is introspection. A reason for thinking that introspection is what Chisholm means by direct knowledge is that he restricts all determiners of justification to conscious states:

> A consequence of our "internalistic" theory of knowledge is that, if one is subject to an epistemic requirement at any time, then this requirement is imposed by the *conscious state* in which one happens to find oneself at that time (*ibid.*, pp. 59-60).

Since he restricts justifiers to conscious states, it is plausible to assume that direct knowledge, for Chisholm, means introspective knowledge, and knowledge by "reflection" coincides with knowledge by introspection.[17] At least in the case of Chisholm, then, $KJ_{dir}$ might be replaced by:

$KJ_{int}$: The only facts that qualify as justifiers of an agent's believing *p* at time *t* are facts that the agent can readily know *by introspection*, at *t*, to obtain or not to obtain.

---

[17] Other epistemologists who restrict justifiers to conscious states or discuss access in terms of introspection include Moser, p. 174; Feldman, "Having Evidence," in D. Austin, ed., *Philosophical Analysis* (Dordrecht: Kluwer, 1988), pp. 83–104; and Audi, "Causalist Internalism," *American Philosophical Quarterly*, XXVI, 4 (1989): 309-20.

Now, the only facts that an agent can know by introspection are facts concerning what conscious states he is (or is not) currently in, so these are the only sorts of facts that qualify as justifiers under $KJ_{mt}$. This form of internalism may be called *strong internalism*:

> SI: Only facts concerning what conscious states an agent is in at time *t* are justifiers of the agent's beliefs at *t*.

Strong internalism, however, is an unacceptable approach to justification, for it has serious, skepticism-breeding, consequences. This is demonstrated by the *problem of stored beliefs*. At any given time, the vast majority of one's beliefs are stored in memory rather than occurrent or active. Beliefs about personal data (for example, one's social security number), about world history, about geography, or about the institutional affiliations of one's professional colleagues, are almost all stored rather than occurrent at a given moment. Furthermore, for almost any of these beliefs, one's conscious state at the time includes nothing that justifies it. No perceptual experience, no conscious memory event, and no premises consciously entertained at the selected moment will be justificationally sufficient for such a belief. According to strong internalism, then, none of these beliefs is justified at that moment. Strong internalism threatens a drastic diminution in the stock of beliefs ordinarily deemed justified, and hence in the stock of knowledge, assuming that justification is necessary for knowledge. This is a major count against this type of theory.

Feldman anticipates this problem because his own account of having evidence also implies that only consciously entertained factors have evidential force (*op. cit.*, pp. 98-99). Feldman tries to meet the threat by distinguishing between occurrent and dispositional senses of epistemic terms. (He actually discusses knowledge rather than justification, but I shall address the issue in terms of justification because that is the target of our investigation.) Feldman is not simply restating the familiar point that 'belief' has occurrent and dispositional senses. He is proposing that the term 'justified' is ambiguous between an occurrent and a dispositional sense. Feldman apparently claims that in the case of stored beliefs, people at most have dispositional justification, not occurrent justification.

There are two problems with this proposal. First, if having a disposition to generate conscious evidential states qualifies as a justifier of a belief, why would this not extend from memorial to perceptual dispositions? Suppose a train passenger awakes from a nap but has not yet opened his eyes. Is he justified in believing propositions about the details of the neighboring landscape? Surely not. Yet he is *dis-*

*posed*, merely by opening his eyes, to generate conscious evidential states that would occurrently justify such beliefs. So the dispositional approach is far too permissive to yield an acceptable sense of 'justified'.[18] Second, can an internalist, especially a strong internalist, live with the idea that certain dispositions count as justifiers? Having or not having a disposition (of the requisite type) is not the sort of fact or condition that can be known by introspection. Thus, the proposal to supplement the occurrent sense of 'justified' with a dispositional sense of 'justified' is simply the abandonment of strong internalism.

## III. INDIRECT KNOWABILITY AND WEAK INTERNALISM

The obvious solution to the problem of stored beliefs is to relax the KJ constraint: allow justifiers to be merely indirectly knowable. This yields:

> $KJ_{ind}$: The only facts that qualify as justifiers of an agent's believing *p* at time *t* are facts that the agent can readily know at *t*, either directly or indirectly, to obtain or not to obtain.

The danger here is that indirect knowledge might let in too much from an internalist perspective. How are externalist forms of knowledge—for example, perceptual knowledge—to be excluded? Clearly, internalism must propose specific forms of knowledge that conform with its spirit. It is fairly clear how internalism should deal with the problem of stored beliefs: simply allow knowledge of justifiers to include memory retrieval. Stored evidence beliefs can qualify as justifiers because the agent can know that they obtain by the compound route of first retrieving them from memory and then introspecting their conscious contents. This yields the following variant of the KJ constraint:

> $KJ_{int+ret}$: The only facts that qualify as justifiers of an agent's believing *p* at time *t* are facts that the agent can readily know, at *t*, to obtain or not to obtain, *by introspection and/or memory retrieval*.

This KJ constraint allows for a more viable form of internalism than strong internalism. We may call it *weak internalism*, and initially articulate it through the following principle:

> WI: Only facts concerning what conscious and/or stored mental states an agent is in at time *t* are justifiers of the agent's beliefs at *t*.

WI will certify the justification of many stored beliefs, because agents often have other stored beliefs that evidentially support them. A person who believes that Washington, D.C. is the capital of the United States may have a stored belief to the effect that a map of the U.S. he recently consulted showed Washington as the capital. The latter stored belief is what justifies the former one. So weak internalism is not plagued with the problem of stored justified beliefs. Weak internalism seems to be a legitimate form of internalism because even stored beliefs qualify, intuitively, as internal states.

---

[18] Feldman might reply that there is an important distinction between memorial and perceptual dispositions; but it is not clear on what basis a principled distinction can be drawn.

Although weak internalism is better than strong internalism, it too faces severe problems. First is the *problem of forgotten evidence*.[19] Many justified beliefs are ones for which an agent once had adequate evidence that she subsequently forgot. At the time of epistemic appraisal, she no longer possesses adequate evidence that is retrievable from memory. Last year, Sally read a story about the health benefits of broccoli in the "Science" section of the *New York Times*. She then justifiably formed a belief in broccoli's beneficial effects. She still retains this belief but no longer recalls her original evidential source (and has never encountered either corroborating or undermining sources). Nonetheless, her broccoli belief is still justified, and, if true, qualifies as a case of knowledge. Presumably, this is because her past acquisition of the belief was epistemically proper. But past acquisition is irrelevant by the lights of internalism (including weak internalism), because only her current mental states are justifiers relevant to her current belief. All past events are "external" and therefore irrelevant according to internalism.

It might be replied that Sally does currently possess evidence in support of her broccoli belief. One of her background beliefs, we may suppose, is that most of what she remembers was learned in an epistemically proper manner. So does she not, after all, now have grounds for the target belief? Admittedly, she has *some* evidence, but is this evidence sufficient for justification? Surely not. In a variant case, suppose that Sally still has the same background belief—namely, that most of what she remembers was learned in an epistemically proper manner—but she in fact acquired her broccoli belief from the *National Inquirer* rather than the *New York Times*. So her broccoli belief was never acquired, or corroborated, in an epistemically sound manner. Then even with the indicated current background belief, Sally cannot be credited with justifiably believing that broccoli is healthful. Her past acquisition is still relevant, and decisive. At least it is relevant so long as we are considering the "epistemizing" sense of justification, in which justification carries a true belief a good distance toward knowledge. Sally's belief in the healthfulness of broccoli is not justified in that sense, for surely she does not know that broccoli is healthful given that the *National Inquirer* was her sole source of information.

---

[19] This sort of problem is discussed by Gilbert Harman, *Change in View* (Cambridge: MIT, 1986); Thomas Senor, "Internalist Foundationalism and the Justification of Memory Belief," *Synthese*, XCIV (1993): 453-76; and Audi, "Memorial Justification," *Philosophical Topics*, XXIII (1995): 31-45.

The category of forgotten evidence is a problem for weak internalism because, like the problem of stored beliefs facing strong internalism, it threatens skeptical outcomes. A large sector of what is ordinarily counted as knowledge are beliefs for which people have forgotten their original evidence.

In reply to the problem of forgotten evidence, Steup[20] offers the following solution. An additional requirement for memorial states to justify a belief that *p*, says Steup, is that the agent have adequate evidence for believing the following counterfactual: "If she had encountered *p* in a questionable source, she would not have formed the belief that *p*." Steup's suggestion is that in the *National Inquirer* variant, Sally fails to have adequate evidence for this counterfactual, and that is why her broccoli belief is not justified. My response to this proposal is twofold. First, the proposed requirement is too strong to impose on memorially justified belief. It is quite difficult to get adequate evidence for the indicated counterfactual. Second, the proposed requirement seems too weak as well. Sally might have adequate evidence for the counterfactual but still be unjustified in holding her broccoli belief. She might have adequate evidence for the counterfactual without its being true; but if it is not true and the rest of the story is as I told it, her broccoli belief is not justified. So Steup's internalist-style solution does not work.

A second problem confronting weak internalism is what I call the *problem of concurrent retrieval*. Principle WI says that *only* conscious and stored mental states are justifiers, but it does not say that *all* sets or conjunctions of such states qualify as justifiers.[21] Presumably, which sets of such states qualify is a matter to be decided by reference to $KJ_{mt+ret}$. If a certain set of stored beliefs can all be concurrently retrieved at time *t* and concurrently introspected, then they would pass the test of $KJ_{mt+ret}$, and could qualify as justifiers under the principle of indirect knowability. But if they cannot all be concurrently retrieved and introspected at *t*, they would fail the test. Now it is clear that the totality of an agent's stored credal corpus at a time cannot be concurrently retrieved from memory. So that set of stored beliefs does not qualify as a justifier for purposes of weak internalism. Unfortunately, this sort of belief set is precisely what certain types of internalist theories require by way of a justifier. Consider holistic coherentism, which says that a proposition *p* is justified for person *S* at time *t* if and only if *p* coheres with *S*'s entire corpus of beliefs at *t* (including, of course, the stored beliefs). A cognitive agent could ascertain, at *t*, whether *p* coheres with her entire corpus only by concurrently retrieving all of her stored beliefs. But such concurrent retrieval is psychologically impossible.[22] Thus, the critically relevant justificational fact under holistic coherentism does not meet even the indirect knowability constraint, much less the direct knowability constraint. Here is a clash, then, between a standard internalist theory of justification and the knowability rationale under scrutiny. Either that rationale is indefensible, or a familiar type of internalism must be abandoned at the outset. Nor is the problem confined to coherentism. Internalist foundationalism might also require concurrent retrieval of more basic (or low-level) beliefs than it is psychologically feasible to retrieve.

---

[20] His proposal was part of his commentary (see acknowledgment note above).

[21] Obviously, one would need to reject the principle that the knowability of fact *A* and the knowability of fact *B* entail the knowability of the conjunctive fact, *A & B*.

## IV LOGICAL AND PROBABILISTIC RELATIONS

As these last examples remind us, every traditional form of internalism involves some appeal to logical relations, probabilistic relations, or their ilk. Foundationalism requires that nonbasically justified beliefs stand in suitable logical or probabilistic relations to basic beliefs; coherentism requires that one's system of beliefs be logically consistent, probabilistically coherent, or the like. None of these logical or probabilistic relations is itself a mental state, either a conscious state or a stored state. So these relations do not qualify as justifiers according to either SI or WI. The point may be illustrated more concretely within a foundationalist perspective. Suppose that Jones possesses a set of basic beliefs at *t* whose contents logically or probabilistically support proposition *p*. This property of Jones's basic beliefs—the property of supporting proposition *p*—is not a justifier under WI, for the property itself is neither a conscious nor a stored mental state. Nor is the possession of this property by these mental states another mental state. So WI has no way of authorizing or permitting Jones to believe *p*. Unless WI is liberalized, no nonbasic belief will be justified, which would again threaten a serious form of skepticism.

---

[22] The "doxastic presumption" invoked by BonJour (*op. cit.*, pp. 101-06) seems to assume that this is possible, but this is simply an undefended assumption. Pollock (*op. cit.*, p. 136) also raises the problem identified here, though in slightly different terms.

Can this problem be remedied by simply adding the proviso that all properties of conscious or stored mental states also qualify as justifiers?[23] This proviso is unacceptably permissive for internalism. One property of many conscious and stored mental states is the property of *being caused by a reliable process*, yet surely internalism cannot admit this archetypically externalist type of property into the class of justifiers. How should the class of properties be restricted? An obvious suggestion is to include only formal properties of mental states, that is, logical and mathematical properties of their contents. But should *all* formal properties be admitted? This approach would fly in the face of the knowability or accessibility constraint, which is the guiding theme of internalism. Only formal properties that are knowable by the agent at the time of doxastic decision should be countenanced as legitimate justifiers under internalism. Such properties, however, cannot be detected by introspection and/or memory retrieval. So some knowing operations suitable for formal properties must be added, yielding a liberalized version of the KJ constraint.

How should a liberalized KJ constraint be designed? The natural move is to add some selected computational operations or algorithms, procedures that would enable an agent to ascertain whether a targeted proposition *p* has appropriate logical or probabilistic relations to the contents of other belief states he is in. Precisely which computational operations are admissible? Again, problems arise. The first is the *problem of the doxastic decision interval*.

The traditional idea behind internalism is that an agent is justified in believing *p at time t* if the evidential beliefs (and perhaps other, nondoxastic states) possessed *at t* have an appropriate logical or probabilistic relation to *p*. In short, justification is conferred simultaneously with evidence possession. Feldman makes this explicit: "For any person S and proposition p and time t, S epistemically ought to believe p at t if and only if p is supported by the evidence S has at t" (*op. cit.*, p. 254). Once the knowability constraint is introduced, however, simultaneous justification looks problematic. If justification is contingent on the agent's ability to know what justifiers obtain, the agent should not be permitted to believe a proposition *p at t* unless she can know *by t* whether the relevant justifiers obtain. Since it necessarily takes some time to compute logical or probabilistic relations, the simultaneity model of justification needs to be revised so that an agent's mental states at *t* justify her in believing only *p* at $t + \epsilon$, for some suitable $\epsilon$. The value of $\epsilon$ cannot be too large, of course, lest the agent's mental states change so as to affect the justificational status of *p*. But $\epsilon$ must be large enough to allow the agent time to determine the relevant formal relations.

---

[23] More precisely, the contemplated proviso should say that the possession of any property by a mental state (or set of mental states) qualifies as a justifier. This reading will be understood wherever the text talks loosely of "properties."

These two conditions—(1) avoid mental change, but (2) allow enough time to compute formal relations—may well be jointly unsatisfiable, which would pose a severe problem for internalism. Mental states, including perceptual states that generate new evidence, change very rapidly and they could easily change before required computations could be executed. On the other hand, although mental states do change rapidly, the agent's belief system might not be epistemically required to reflect or respond to each change until interval $\epsilon$ has elapsed. Some doxastic decision interval, then, might be feasible.

Is there a short enough decision interval during which justificationally pertinent formal properties can be computed? Coherentism says that *S* is justified in believing proposition *p* only if *p* coheres with the rest of *S*'s belief system held at the time. Assume that coherence implies logical consistency. Then coherentism requires that the logical consistency or inconsistency of any proposition *p* with *S*'s belief system must qualify as a justifier. But how quickly can consistency or inconsistency be ascertained by mental computation? As Christopher Cherniak[24] points out, determination of even tautological consistency is a computationally complex task in the general case. Using the truth-table method to check for the consistency of a belief system with 138 independent atomic propositions, even an ideal computer working at "top speed" (checking each row of a truth table in the time it takes a light ray to traverse the diameter of a proton) would take twenty billion years, the estimated time from the "big-bang" dawn of the universe to the present. Presumably, twenty billion years is not an acceptable doxastic decision interval!

Any reasonable interval, then, is too constraining for garden-variety coherentism. The knowability constraint again clashes with one of the stock brands of internalism.[25] Dyed-in-the-wool internalists might be prepared to live with this result. "So much the worse for traditional coherentism," they might say, "we can live with its demise." But this does not get internalism entirely off the hook. There threaten to be many logical and probabilistic facts that do not qualify as justifiers because they require too long a doxastic interval to compute. Furthermore, it is unclear what is a principled basis for deciding what is too long. This quandary confronting internalism has apparently escaped its proponents' attention.

---

[24] "Computational Complexity and the Universal Acceptance of Logic," this JOURNAL, LXXXI, 12 (December 1984): 739-58.

[25] This computational difficulty for coherentism is identified by Kornblith, "The Unattainability of Coherence," in J. Bender, ed., *The Current State of the Coherence Theory* (Dordrecht: Kluwer, 1989), pp. 207–14.

A second problem for logical and probabilistic justifiers is the *availability problem*. Suppose that a particular set of *computational operations*—call it COMP—is provisionally selected for inclusion alongside introspection and memory retrieval. COMP might include, for example, a restricted (and hence noneffective) use of the truth-table method, restricted so as to keep its use within the chosen doxastic decision interval.[26] This yields a new version of the KJ constraint:

> $KJ_{int+ret+COMP}$: The only facts that qualify as justifiers of an agent's believing $p$ at time $t$ are facts that the agent can readily know within a suitable doxastic decision interval *via introspection, memory retrieval, and/or COMP.*

Now, the KJ constraint is presumably intended to apply not only to the cleverest or best-trained epistemic agents but to all epistemic agents, including the most naive and uneducated persons on the street. After all, the point of the knowability constraint is that justifiers should be facts within the purview of every epistemic agent. Under the GD conception, compliance with epistemic duty or responsibility is not intended to be the private preserve of the logical or mathematical elite. It is something that ought to be attained—and should therefore be attainable—by any human agent. The truth-table method, however, does not seem to be in the intellectual repertoire of naive agents, so it is illegitimate to include COMP operations within a KJ constraint. Unlike introspection and memory retrieval, it is not available to all cognitive agents.

It may be replied that computational operations of the contemplated sort would be within the *capacity* of normal human agents. No super-human computational powers are required. Computing power, however, is not the issue. A relevant sequence of operations must also be *available* in the agent's intellectual repertoire; that is, she must know which operations are appropriate to obtain an answer to the relevant (formal) question.[27] Since truth-table methods and other such algorithms are probably not in the repertoire of ordinary cognitive agents, they cannot properly be included in a KJ constraint.

---

[26] Because of the contemplated restriction, there will be many questions about formal facts to which COMP cannot deliver answers. Thus, formal facts that might otherwise qualify as justifiers will not so qualify under the version of the KJ constraint that incorporates COMP.

[27] Propositional (or "declarative") knowledge of the appropriate sequence of operations is, perhaps, an unduly restrictive requirement. It would suffice for the agent to have "procedural" skills of the right sort. But even such skills will be lacking in naive cognitive agents.

A third problem concerns the proper methodology that should be used in selecting a KJ constraint that incorporates computational operations. As we see from the first two problems, a KJ constraint that conforms to the spirit of the GD rationale must reflect the basic cognitive skills or repertoires of actual human beings. What these basic repertoires consist in, however, cannot be determined a priori. It can only be determined with the help of empirical science. This fact fundamentally undermines the methodological posture of internalism, a subject to which I shall return in section VII.

Until now, I have assumed a *universal* accessibility constraint, one that holds for all cognitive agents. But perhaps potential justifiers for one agent need not be potential justifiers for another. Justifiers might be allowed to vary from agent to agent, depending on what is knowable by the particular agent. If two agents have different logical or probabilistic skills, then some properties that do not qualify as justifiers for one might yet qualify as justifiers for the other. Indeed, the constraint $KJ_{\mathrm{int+ret+COMP}}$ might be read in precisely this agent-relativized way. The subscripts may be interpreted as indicating knowledge routes that are available *to the agent in question*, not necessarily to all agents.

If KJ constraints are agent relativized as a function of differences in knowledge skills, this means that two people in precisely the same evidential state (in terms of perceptual situation, background beliefs, and so on) might have different epistemic entitlements. But if the two agents are to comply with their respective epistemic duties, each must *know* which knowledge skills she has. This simply parallels the second step of the internalist's original three-step argument. If one's epistemic duties or entitlements depend on one's knowledge skills (for example, on one's computational skills), then compliance with one's duties requires knowledge of which skills one possesses. There are two problems with this approach. First, it is unlikely that many people—especially ordinary people on the street—have this sort of knowledge, and this again threatens large-scale skepticism. Second, what is now required to be known by the agent is something about the *truth-getting* power of her cognitive skills—that is, the power of her skills in detecting justifiers. This seems to be precisely the sort of *external* property that internalists regard as anathema. How can they accept this solution while remaining faithful to the spirit of internalism?[28]

---

[28] It might be argued that internalism's spirit leads to a similar requirement even for universal versions of a KJ constraint, not just for agent-relativized versions. Perhaps so; but so much the worse for the general form of internalism.

### V. EPISTEMIC PRINCIPLES

When the KJ constraint speaks of justifiers, it is not clear exactly what these comprehend. Specifically, do justifiers include epistemic principles themselves? I believe that principles should be included, because epistemic principles are among the items that determine whether or not an agent is justified in believing a proposition, which is just how 'justifiers' was defined. Furthermore, true epistemic principles are items an agent must know if she is going to determine her epistemic duties correctly. Knowledge of her current states of mind and their properties will not instruct her about her epistemic duties and entitlements unless she also knows true epistemic principles.

How are epistemic principles to be known, according to internalism? Chisholm[29] says that central epistemic principles are normative supervenience principles, which (when true) are necessarily true. Since they are necessary truths, they can be known a priori—in particular, they can be known "by reflection."

> The internalist assumes that, merely by reflecting upon his own conscious state, he can formulate a set of epistemic principles that will enable him to find out, with respect to any possible belief he has, whether he is justified in having that belief.[30]

This passage is ambiguous as to whether (correct) epistemic principles are accessible on reflection just to epistemologists, or accessible to naive epistemic agents as well. The latter, however, must be required by internalism, because justifiers are supposed to be determinable by all epistemic agents.

Are ordinary or naive agents really capable of formulating and recognizing correct epistemic principles? This seems highly dubious. Even many career-long epistemologists have failed to articulate and appreciate correct epistemic principles. Since different epistemologists offer disparate and mutually conflicting candidates for epistemic principles, at most a fraction of these epistemologists can be right. Perhaps none of the principles thus far tendered by epistemologists is correct! In light of this shaky and possibly dismal record by professional epistemologists, how can we expect ordinary people, who are entirely ignorant of epistemology and its multiple pitfalls, to succeed at this task?[31] Nor is it plausible that they should succeed at this task purely "by reflection" on their conscious states, since among the matters epistemic principles must resolve is what computational skills are within the competence of ordinary cognizers. I do not see how this can be answered a priori, "by reflection."

---

[29] "The Status of Epistemic Principles," *Noûs*, xxiv (1990): 209-15.

[30] *Theory of Knowledge*, 3rd edition, p 76; emphasis omitted.

[31] A similar worry is expressed by Alston in "Internalism and Externalism in Epistemology," pp. 221-22.

A crippling problem emerges for internalism. If epistemic principles are not knowable by all naive agents, no such principles can qualify as justifiers under the KJ constraint. If no epistemic principles so qualify, no proposition can be justifiably believed by any agent. Wholesale skepticism follows.

## VI THE CORE DILEMMA FOR THE THREE-STEP ARGUMENT

I raise doubts here about whether there is any cogent inferential route from the GD conception to internalism via an acceptable KJ constraint. Here is the core dilemma. The minimal, unvarnished version of the KJ constraint does not rationalize internalism. That simple constraint merely says that justifiers must be readily knowable, and some readily knowable facts might be external rather than internal. If *all* routes to knowledge of justifiers are allowed, then knowledge by perception must be allowed. If knowledge by perception is allowed, then facts of an external sort could qualify for the status of justifiers. Of course, no epistemologist claims that purely external facts should serve as justifiers. But partly external facts are nominated by externalists for the rank of justifiers. Consider properties of the form: being a reliable perceptual indicator of a certain environmental fact. This sort of property is at least partly external because reliability involves truth, and truth (on the usual assumption) is external. Now suppose that a certain auditory perceptual state has the property of being a reliable indicator of the presence of a mourning dove in one's environment. Might the possession of this reliable indicatorship property qualify as a justifier on the grounds that it is indeed readily knowable? If every route to knowledge is legitimate, I do not see how this possibility can be excluded. After all, one could use past perceptions of mourning doves and their songs to determine that the designated auditory state is a reliable indicator of a mourning dove's presence. So if unrestricted knowledge is allowed, the (partly) external fact in question might be perfectly knowable. Thus, the unvarnished version of the KJ constraint does not exclude external facts from the ranks of the justifiers.

The simple version of the KJ constraint, then, does not support internalism. Tacit recognition of this is what undoubtedly leads internalists to favor a "direct" knowability constraint. Unfortunately, this extra rider is not rationalized by the GD conception. The GD conception at best implies that cognitive agents must know what justifiers are present or absent. No particular *types* of knowledge, or *paths* to knowledge, are intimated. So the GD conception cannot rationalize a restrictive version of the KJ constraint that unambiguously yields internalism.

Let me put the point another way. The GD conception implies that justifiers must be readily knowable, but are internal facts always *more readily* knowable than external facts? As discussed earlier, probabilistic relations presumably qualify as internal, but they do not seem to be readily knowable by human beings. An entire tradition of psychological research on "biases and heuristics" suggests that naive agents commonly commit probabilistic fallacies, such as the "conjunction fallacy," and use formally incorrect judgmental heuristics, such as the representativeness heuristic and the anchoring-and-adjustment heuristic.[32] If this is right, people's abilities at detecting probabilistic relationships are actually rather weak. People's perceptual capacities to detect external facts seem, by contrast, far superior. The unqualified version of the KJ constraint, therefore, holds little promise for restricting all justifiers to internal conditions in preference to external conditions, as internalism requires.[33]

## VII THE METHODOLOGY OF EPISTEMOLOGY: EMPIRICAL OR A PRIORI?

Internalism standardly incorporates the doctrine that epistemology is a purely a priori or armchair enterprise rather than one that needs help from empirical science. Chisholm puts the point this way:

> The epistemic principles that [the epistemologist] formulates are principles that one may come upon and apply merely by sitting in one's armchair, so to speak, and without calling for any outside assistance. In a word, one need only consider one's own state of mind.[34]

Previous sections already raised doubts about the merits of apriorism in epistemology, even in the context of the theoretical architecture presented here. I now want to challenge the viability of apriorism in greater depth.

Assume that, despite my earlier reservations, an internalist restriction on justifiers has somehow been derived, one that allows only conscious states and certain of their nonexternal properties to serve as justifiers. How should the epistemologist identify particular conscious states and properties as justifiers for specific propositions (or types of propositions)? In other words, how should specific epistemic principles be crafted? Should the task be executed purely a priori, or can scientific psychology help?

---

[32] See Amos Tversky and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases," in Kahneman, P. Slovic, and Tversky, eds., *Judgment under Uncertainty* (New York: Cambridge, 1982), pp. 3–20; and Tversky and Kahneman, "Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment," *Psychological Review*, XC (1983) 293-315.

[33] It is not really clear, moreover, why logical or probabilistic facts intuitively count as "internal" facts. They certainly are not internal in the same sense in which mental states are internal. This is an additional problem about the contours of internalism.

[34] *Theory of Knowledge*, 3rd edition, p 76.

For concreteness, consider justifiers for memory beliefs. Suppose an adult consciously remembers seeing, as a teenager, a certain matinee idol. This ostensible memory could have arisen from imagination, since he frequently fantasized about this matinee idol and imagined seeing her in person. What clues are present in the current memory impression by which he can tell whether or not the recollection is veridical? This is precisely the kind of issue which internalist epistemic principles should address. If there are no differences in features of memory states that stem from perceptions of real occurrences versus features of states that stem from mere imagination, does this not raise a specter of skepticism over the domain of memory? If there are no indications by which to distinguish veridical from nonveridical memory impressions, can we be justified in trusting our memory impressions? Skepticism aside, epistemologists should surely be interested in identifying the features of conscious memory impressions by which people are made more or less justified (or prima facie justified) in believing things about the past.

Epistemologists have said very little on this subject. Their discussions tend to be exhausted by characterizations of memory impressions as "vivid" or "nonvivid." There is, I suspect, a straightforward reason for the paucity of detail. It is extremely difficult, using purely armchair methods, to dissect the microfeatures of memory experiences so as to identify telltale differences between trustworthy and questionable memories. On the other hand, empirical methods have produced some interesting findings, which might properly be infused into epistemic principles in a way entirely congenial to internalism. Important research in this area has been done by Marcia Johnson and her colleagues.[35] I shall illustrate my points by brief reference to their research.

Johnson calls the subject of some of her research *reality monitoring*. She tries to characterize the detectable differences between (conscious) memory traces derived from veridical perception of events versus memory traces generated by mere imaginations of events.[36] Johnson and Raye (*op. cit.*) propose four dimensions along which memory cues will typically differ depending on whether their origin was perceptual or imaginative. As compared with memories that originate from imagination, memories originating from perception tend to have (1) more perceptual information (for example, color and sound), (2) more contextual information about time and place, and (3) more meaningful detail. When a memory trace is rich along these three dimensions, this is evidence of its having originated through perception. Memories originating from imagination or thought, by contrast, tend to be rich on another dimension: they contain more information about the cognitive operations involved in the original thinkings or imaginings (for example, effortful attention, image creation, or search). Perception is more automatic than imagination, so a memory trace that originates from perception will tend to lack attributes concerning effortful operations. Johnson and Raye therefore suggest that differences in average value along these types of dimensions can form the basis for deciding whether the origin of a memory is perceptual or nonperceptual. A memory with a great deal of visual and spatial detail, and without records of intentional constructive and organizational processes, should be judged to have been perceptually derived.[37]

---

[35] See Johnson and Carol Raye, "Reality Monitoring," *Psychological Review*, LXXXVIII (1981): 67-85; and Johnson, Mary Foley, Aurora Suengas, and Raye, "Phenomenal Characteristics of Memories for Perceived and Imagined Autobiographical Events," *Journal of Experimental Psychology: General*, CXVII (1988): 371-76.

Epistemologists would be well-advised to borrow these sorts of ideas and incorporate them into their epistemic principles. A person is (prima facie) justified in believing in the real occurrence of an ostensibly recalled event if the memory trace is strong on the first three dimensions and weak on the fourth dimension. Conversely, an agent is unjustified in believing in the real occurrence of the recalled event if the memory trace is strong on the fourth dimension but weak on the first three dimensions. All of these dimensions, of course, concern features of conscious experience. For this reason, internalist epistemologists should be happy to incorporate these kinds of features into their epistemic principles.

Let me distinguish two categories of epistemologically significant facts about memory experience which empirical psychology might provide. First, as we have seen, it might identify types of representational materials which are generally available in people's memory experiences. Second, it might indicate which of these representational materials are either reliable or counter-reliable indicators of the veridicality of the ostensibly recalled events. Is the reliability of a memory cue a legitimate issue from an internalist perspective? It might be thought not, since reliability is usually classed as an external property. But epistemologists might use reliability considerations to decide which memory characteristics should be featured in epistemic principles. They need not insert reliability per se into the principles. There is nothing in our present formulation of internalism, at any rate, which bars the latter approach. Any KJ constraint provides only a necessary condition for being a justifier; it leaves open the possibility that additional necessary conditions, such as reliable indication, must also be met. Indeed, many internalists do use reliability as a (partial) basis for their choice of justifiers. BonJour (*op. cit.*, p. 7) says that the basic role of justification is that of a *means* to truth, and he defends coherence as a justifier on the ground that a coherent system of beliefs is likely to correspond to reality. This point need not be settled definitively, however. There are already adequate grounds for claiming that internalism cannot be optimally pursued without help from empirical psychology, whether or not reliability is a relevant consideration.

---

[36] Memory errors are not confined, of course, to confusions of actual with imagined events. There are also errors that arise from confusing, or blending, two actual events. But this research of Johnson's focuses on the actual/nonactual (or perceived versus imagined) problem.

[37] They also recognize that people can compare a target memory with memories of contextually related events to assess the target's veridicality. This kind of "coherence" factor is a stock-in-trade of epistemology, however, and hence not a good example of the distinctive contributions psychology can make to this subject. I therefore pass over it.

## VIII CONCLUSION

Let us review the parade of problems infecting internalism which we have witnessed, though not all in their order of presentation. (1) The argument from the GD conception of justification to internalism does not work. Internalism can be derived only from a suitably qualified version of the KJ constraint because the unqualified version threatens to allow external facts to count as justifiers. No suitably qualified version of the KJ constraint is derivable from the GD conception. (2) A variety of qualified KJ constraints are possible, each leading to a different version of internalism. None of these versions is intuitively acceptable. Strong internalism, which restricts justifiers to conscious states, is stuck with the problem of stored beliefs. Weak internalism, which allows stored as well as conscious beliefs to count as justifiers, faces the problem of forgotten evidence and the problem of concurrent retrieval. (3) The question of how logical and probabilistic facts are to be included in the class of justifiers is plagued by puzzles, especially the puzzle of the doxastic decision interval and the issue of availability. (4) Epistemic principles must be among the class of justifiers, but such principles fail internalism's knowability requirement. (5) The favored methodology of internalism—the armchair method—cannot be sustained even if we grant the assumption that justifiers must be conscious states.

Internalism is rife with problems. Are they all traceable to the GD rationale? Could internalism be salvaged by switching to a different rationale? A different rationale might help, but most of the problems raised here arise from the knowability constraint. It is unclear exactly which knowability constraint should be associated with internalism, and all of the available candidates generate problematic theories. So I see no hope for internalism; it does not survive the glare of the spotlight.

ALVIN I. GOLDMAN

University of Arizona

# CALIBRATION AND THE EPISTEMOLOGICAL ROLE OF BAYESIAN CONDITIONALIZATION *

Suppose that a scientist acquires some new evidence; she learns that $e$ obtains. She must update her confidence in hypothesis $h$ accordingly. *Bayesian conditionalization* (BC) holds that her new confidence in $h$ should equal the confidence that she had (prior to discovering $e$) that $h$ obtains in the event that $e$ obtains. More precisely, if $\mathrm{pr}_{\mathrm{old}}(\cdot)$ is our probability distribution before discovering $e$, and $\mathrm{pr}_{\mathrm{new}}(\cdot)$ is our probability distribution after discovering $e$, then BC says that:

> *If* $0 < \mathrm{pr}_{\mathrm{old}}(h) < 1$ (so that before discovering $e$, we regarded $h$ as a live hypothesis, subject to confirmation and disconfirmation)
> $\mathrm{pr}_{\mathrm{old}}(e) < 1$ (so that $e$ was not believed already)
> $e$ is the strongest claim whose truth we take ourselves to have discovered
> and $\mathrm{pr}_{\mathrm{old}}(e) > 0$ (so that $\mathrm{pr}_{\mathrm{old}}(h|e) = \mathrm{pr}_{\mathrm{old}}(h \,\&\, e)/\mathrm{pr}_{\mathrm{old}}(e)$)
> *then* $\mathrm{pr}_{\mathrm{new}}(h) = \mathrm{pr}_{\mathrm{old}}(h|e)$.

Coupled with Bayes's theorem—$\mathrm{pr}_{\mathrm{old}}(h|e) = \mathrm{pr}_{\mathrm{old}}(h)\ \mathrm{pr}_{\mathrm{old}}(e|h)/\mathrm{pr}_{\mathrm{old}}(e)$—BC seems able to account elegantly for many important features of theory confirmation in science, such as the value of diverse evidence, the kernel of correctness in hypothetico-deductivism, the roles of auxiliary hypotheses and rival hypotheses and background beliefs, and the failure in certain cases of hypotheses to be confirmed by their positive instances.[1]

In section I, I review certain considerations often mentioned as posing difficulties for BC. Curiously, several of these purported challenges involve cases that, strictly speaking, fall outside BC's jurisdiction. Why, then, are these considerations widely perceived as making trouble for BC? At least partly because (I suggest) we expect BC to capture not merely our intuitions regarding how we should *revise* our opinions (that is, update our degrees of confidence in various claims) upon discovering new evidence, but also our intuitions regarding whether *e counts as evidence for h*; we tend to regard BC as purporting to explain why vari-

---

* My thanks to Jeff Barrett, Jim Bogen, Mitch Green, Alan Hajek, Michael Otsuka, Ric Otte, Jim Woodward, and Lyle Zynda for discussions of early drafts, and to Bill Talbott for more recent discussion.

[1] Many authors have argued that BC can account for these features. For discussion, see John Earman, *Bayes or Bust?* (Cambridge: MIT, 1992), chapter 3 ("Success Stories"). A careful examination of these arguments and their potential defects is beyond the scope of this paper.

0022-362X/99/9606/294–324

ous factors make *e* stronger or weaker evidence for or against *h*. But (as we shall see) the notion of *e*'s being 'evidence for *h*' is not restricted to cases where we discover *e*, $0 < \mathrm{pr}_{\mathrm{old}}(e) < 1$, $0 < \mathrm{pr}_{\mathrm{old}}(h) < 1$, and we update our confidence in *h* accordingly. So even if BC governs every such updating, it does not thereby automatically account for every case in which *e* constitutes evidence for (or against) *h*.

In section II, I distinguish two roles that BC might play in an account of confirmation: as governing how we should *update* our opinions upon discovering *new* evidence, and as governing the steps in the *arguments* by which our current opinions are to be justified by the evidence that we have *already* assembled. The former is a diachronic matter; the latter concerns the arguments that we give at a particular moment for our opinions then. Although this is an elementary distinction, I have not seen it drawn elsewhere. This distinction is useful, I argue, because if BC governs justificatory arguments, then it is in the right position to cash out what it is for *e* to be evidence for (or against) *h*. In this role, BC could account for the various features of confirmation that it seems able to explicate so nicely, and yet avoid the problems that result from trying to explicate '*e* is evidence for *h*' in terms of the difference that *e*'s discovery makes, made, or would have made to our confidence in *h*.[1]

In section III, I consider *why* it would be wrong for a justificatory argument to violate BC–what sort of logical error would thereby be made. Diachronic Dutch Book arguments have sometimes been thought to explicate why it would be wrong for us to violate BC in updating our opinions. But these arguments have encountered formidable obstacles. Indeed, for reasons I discuss in section I, I think that it is *rationally permissible* for an *update* to violate BC. I use the notion of *calibration* to elaborate the sense in which it would be wrong for a *justificatory argument* to violate BC. I demonstrate that a scientist's opinions at two steps of a justificatory argument can all be calibrated (in the relevant sense) only if her transition from one step to the next accords with BC. This calibration argument, I contend, avoids difficulties encountered by diachronic Dutch Book arguments.

In applying BC to justificatory arguments rather than to diachronic updates, I aim to find a role for BC to play where it could account for various important features of confirmation.

---

[1] I do not contend that all serious obstacles to reconstructing scientific practice in accord with BC can be avoided in this way. In particular, logical omniscience (which I assume throughout) and the problem of the priors require separate treatment. Also, I presume throughout that upon making an observation, we accept a claim whose content expresses our new evidence, and that any evidence presented in a justificatory argument consists of a proposition whose truth has been ascertained.

## I

I shall begin by briefly reviewing four alleged problems for BC and trying to understand just why they are supposed to be problems.

First, consider the *problem of old evidence*. Suppose that $h$ was formulated only *after* $e$ was discovered. Since we already believe $e$ by the time we formulate $h$, our $\mathrm{pr}_{\mathrm{old}}(e)$ at that time is 1, so $\mathrm{pr}_{\mathrm{old}}(e|h) = 1$, so (by Bayes's theorem) $\mathrm{pr}_{\mathrm{old}}(h|e) = \mathrm{pr}_{\mathrm{old}}(h)$, and so (by BC) $\mathrm{pr}_{\mathrm{new}}(h) = \mathrm{pr}_{\mathrm{old}}(h)$. Yet we may well regard $e$ as evidence for (or against) $h$.

So the problem of old evidence is typically posed.[3] But BC applies only when, having already formulated $h$, we discover $e$ and must update our confidence in $h$ accordingly; recall that BC is specifically restricted to cases where $\mathrm{pr}_{\mathrm{old}}(e) < 1$. When $e$ is old evidence, there is no moment at which $e$'s discovery elicits a shift in our confidence in $h$, since there is no $\mathrm{pr}_{\mathrm{old}}(h)$, prior to discovering $e$, from which to shift. Old evidence therefore falls outside of BC's official scope. This makes it unclear why old evidence is commonly perceived as posing a problem for BC.

I suggest that one reason why old evidence is taken to be a problem for BC is because BC is intended to account for the various considerations that may make $e$ evidence for $h$. Old evidence is a problem for this project because intuitively, old evidence can be *evidence for h*, even though old evidence cannot provoke our *updating* our confidence in $h$, and so BC apparently tells us nothing about what makes some old evidence $e$ stronger or weaker evidence for or against $h$. In other words, although BC applies only to updatings, Bayesians purport to be giving an analysis of what it is for evidence to "count for or against a theory, or be neutral towards it," as Colin Howson and Peter Urbach[4] put it; it is quite common to treat '$e$ is evidence for $h$' as one way of expressing the target of confirmation theory.[5] But no account–Bayesian or otherwise–that applies only to updatings can *straightforwardly* explain when $e$ counts as evidence for $h$, since not all evidence for $h$ provokes, upon its discovery, an increase in our confidence in $h$–as when $e$ is discovered before we have even formulated $h$. Thus, old evidence is a problem for BC.

One way to make an account of updating explain when old evidence $e$ counts as evidence for $h$, even though our confidence in $h$ is not *actually* updated upon our discovering $e$, is to apply the account of updating to a *counterfactual* updating: how we should have shifted our confidence in $h$ upon discovering $e$, *had* we formulated $h$ before discovering $e$. This appeal to counterfactual updatings has been extensively criticized, though not all find these criticisms persuasive.[6] I cannot now give this proposal the scrutiny it deserves, but I need to mention one of its chief difficulties. What our $\mathrm{pr}_{\mathrm{old}}(h)$ would have been, had we formulated $h$ sometime before discovering $e$, depends upon the background beliefs that we held before discovering $e$ regarding auxiliary hypotheses and alternatives to $h$. These background beliefs may be quite foreign to us when we judge, long after $e$ was discovered, whether $e$ is evidence for $h$. Consequently, such historical considerations seem out of place–intuitively irrelevant to $e$'s bearing on $h$.

---

[3] For a good review of the literature, see Earman, pp. 119-36.

[4] *Scientific Reasoning: The Bayesian Approach* (Chicago: Open Court, 1993, 2nd edition), p. 117

[5] See, for instance, Paul Horwich, *Probability and Evidence* (New York: Cambridge, 1982), p. 51; and Mark Kaplan, *Decision Theory as Philosophy* (New York: Cambridge, 1996), p. 48.

Of course, we might avoid this problem by applying BC not to the probability distribution that we would have held prior to discovering $e$, had $h$ already been formulated by then, but to some other probability distribution, which is not picked out by some exercise in counterfactual history. Presumably, by proper selection of this reference probability distribution, BC could be made to yield verdicts that agree with our intuitions. But we might well wonder why $e$'s status now as evidence for $h$ depends upon the way that we should have updated our opinions upon discovering $e$, were we beginning from a certain probability distribution that is neither our actual distribution at some past moment nor even the distribution that we would have held then, had $h$ been formulated by then. I shall later suggest that the reference probability distribution is relevant because it appears in the argument by which we justify our current degree of confidence in $h$. Therefore, whether $e$ is evidence for $h$ depends not on some hypothetical updating, but on our current justificatory arguments.

If we cash out '$e$ is evidence for $h$' in terms of some actual or counterfactual updating, depending upon whether $e$ is new or old evidence, then we run the risk of construing '$e$ is evidence for $h$' very differently in these two cases. Intuitively, however, the temporal order should not enter into it. I shall ultimately give '$e$ is evidence for $h$' a uniform treatment, whatever the temporal order, by elaborating it in terms of $e$'s role in an argument justifying our current degree of confidence in $h$. Actually, even when $e$ was discovered *after* $h$ was formulated but nevertheless some time ago, we should not construe '$e$ is evidence for $h$' in terms of the actual updating that was then provoked by $e$'s discovery.[7] Whether $e$

---

[6] For critique, see Earman, pp. 123, 134, and the references cited there For defense, see, for example, Howson and Urbach, pp. 403-08.

[7] Ellery Eells terms this $e$ 'old new evidence' in his "Problems of Old Evidence," *Pacific Philosophical Quarterly*, LXVI (1985). 283-302.

*now* counts as evidence for *h* does not depend on whether *at the time it was discovered*, *e* provoked an updating that raised our confidence in *h*. That past updating reflected our opinions *then* regarding various auxiliary and rival hypotheses, whereas it is our *current* opinions that influence whether *e now* counts as evidence for *h*. (Of course, whether *e* now counts as evidence for *h* cannot be fixed by whether, right now, $\text{pr}(h|e) > \text{pr}(h)$, since *e* was discovered some time ago, so these two probabilities are equal.) Again, I shall suggest that '*e* is evidence for *h*' should be understood in terms of *e*'s role in arguments by which we now justify our current degree of confidence in *h*.

The second of the four alleged problems for BC that I wish to discuss is that, if *h* is already believed (that is, $\text{pr}_{\text{old}}(h) = 1$), then discovery of some new evidence *e* cannot confirm *h* (that is, cannot raise our confidence in *h*) under BC (or any other updating rule that respects the axioms of probability). Likewise, *e* cannot disconfirm *h* if $\neg h$ is already believed (that is, $\text{pr}_{\text{old}}(h) = 0$). Again, although this is reputed[8] to be a problem for BC, it is initially unclear just why. Like old evidence, this case falls outside BC's jurisdiction, since BC requires $0 < \text{pr}_{\text{old}}(h) < 1$. One reason it poses a problem for BC, I suggest, is again that we expect BC to provide an analysis of '*e* is evidence for *h*', and this project is threatened when *e* intuitively supports (or is evidence against) *h* and yet *e*'s discovery fails to result in our updating our opinion of *h*. Once again, I shall shortly suggest that the fault lies not with BC, but with using it to cash out '*e* is evidence for *h*' in terms of updating rather than in terms of an argument justifying our current opinion of *h*.

The third alleged problem for BC is a sort of dual to the second: if *h* is already believed (that is, $\text{pr}_{\text{old}}(h) = 1$), then according to BC, *e*'s discovery cannot disconfirm *h*—cannot lower our confidence in *h*—and if $\neg h$ is already believed (that is, $\text{pr}_{\text{old}}(h) = 0$), then according to BC, *e* cannot confirm *h*. But this is inconsistent with scientific practice; new evidence sometimes leads scientists to take seriously a theory that they had rejected or to challenge a theory that they had accepted.

Again, however, it is not initially apparent why these considerations make any trouble for BC. As Isaac Levi[9] has noted, BC purports to cover only shifts of opinion upon discovering some *e that is consistent with what we already fully believe*. If $\text{pr}_{\text{old}}(e) = 0$, as when *h* logically entails $\neg e$ and $\text{pr}_{\text{old}}(h) = 1$, then our rejection of *h* upon discovering *e* is entirely consistent with BC, just outside BC's scope. For that matter, BC requires $0 < \text{pr}_{\text{old}}(h) < 1$, so any rise of suspicions regarding previously accepted hypotheses or rehabilitation of previously rejected hypotheses goes beyond BC's jurisdiction. Such shifts of opinion are consistent with BC.

---

[8] See, for instance, Kaplan, p. 50.

[9] *The Enterprise of Knowledge* (Cambridge: MIT, 1980), p. 82.

One reason, I think, that such shifts of opinion are often perceived to pose a problem for BC is again that BC is expected to explain why $e$ is strong or weak evidence for or against $h$. Intuitively, the factors relevant to $e$'s evidential power are the same whether $h$ is already fully believed, already rejected, or being entertained as a live hypothesis. But an account of '$e$ is evidence for $h$' which uses BC to cover the cases where $h$ is a live hypothesis, and some other principles for the other cases, threatens to treat these cases in a counterintuitively nonuniform way. We might well be prepared to grant that along with BC to cover shifts of opinion upon consistent extension of our full beliefs, there are further rules covering *updates* that involve rehabilitating formerly rejected hypotheses or challenging formerly accepted hypotheses.[10] These shifts of opinion are intuitively quite different from "business as usual." But we may be much less willing to grant that an explication of '$e$ is evidence for $h$' should treat these cases differently. Intuitively, the same considerations that render $e$ powerful evidence against $h$, when $0 < \text{pr}_{\text{old}}(h) < 1$, should have made $e$ powerful evidence against $h$ even if we had already been taking $\text{pr}_{\text{old}}(h) = 1$. BC's limited range of application thus becomes an argument against its appearing in an analysis of '$e$ is evidence for $h$'.

Of course, we might try to address this problem by expanding BC's scope. We might begin by making $\text{pr}_{\text{old}}(h|e)$ well defined even when $\text{pr}_{\text{old}}(e) = 0$, as Bruno de Finetti[11] does by holding $\text{pr}_{\text{old}}(h|e)$ to be unconstrained by unconditional probabilities when $\text{pr}_{\text{old}}(e) = 0$; $\text{pr}_{\text{old}}(h|e)$ is then no longer equal to $\text{pr}_{\text{old}}(h \,\&\, e)/\text{pr}_{\text{old}}(e)$ or to $\text{pr}_{\text{old}}(h)\text{pr}_{\text{old}}(e|h)/\text{pr}_{\text{old}}(e)$. We could then hold that even when $\text{pr}_{\text{old}}(e) = 0$, $e$'s evidential relevance to $h$ is given by BC–by comparing $\text{pr}_{\text{old}}(h|e)$ to $\text{pr}_{\text{old}}(h)$. But this move alone cannot resolve the problem. To explain why $e$ is strong or weak evidence for or against $h$, we must still give some account of why $\text{pr}_{\text{old}}(h|e)$ takes on the value it does. Since this account fails to link $\text{pr}_{\text{old}}(h|e)$ to $\text{pr}_{\text{old}}(h)$, $\text{pr}_{\text{old}}(e)$, and the like in the usual way, the explanation of $e$'s evidential relevance to $h$ will be quite different from BC's explanation in cases where $0 < \text{pr}_{\text{old}}(h) < 1$ and $\text{pr}_{\text{old}}(e) > 0$. For example, BC is often cited as explaining what is right about hypothetico-deductivism–why the fact that $h$ logically entails $e$ can sometimes make $e$ evidence for $h$: if $h$ logically entails $e$, then (when $\text{pr}_{\text{old}}(h) > 0$) $\text{pr}_{\text{old}}(e|h) = 1$, and so $\text{pr}_{\text{new}}(h) =$

---

[10] For example, "contraction strategies" are discussed by Levi, pp. 59-62.

[11] *Probability, Induction, and Statistics* (New York: Wiley, 1972), pp. 81-82. See also, for instance, Levi, p. 219; and Earman, p. 36.

$\mathrm{pr}_{\mathrm{old}}(h|e) = \mathrm{pr}_{\mathrm{old}}(h)/\mathrm{pr}_{\mathrm{old}}(e)$, which (when $\mathrm{pr}_{\mathrm{old}}(e) < 1$) exceeds $\mathrm{pr}_{\mathrm{old}}(h)$. But this argument presumes $\mathrm{pr}_{\mathrm{old}}(e) > 0$; if $\mathrm{pr}_{\mathrm{old}}(e) = 0$, we would need an utterly different account of why $\mathrm{pr}_{\mathrm{old}}(h|e) > \mathrm{pr}_{\mathrm{old}}(h)$.

The fourth alleged problem for BC that I shall raise–and, arguably, the most fundamental–I shall call the *problem of incorrigibility*. If a scientist obeys BC after discovering $e$, and then again after discovering $e'$, her $\mathrm{pr}_{\mathrm{new}}(h)$ is her $\mathrm{pr}_{\mathrm{old}}(h|e \,\&\, e')$. Therefore, a scientist who with each new observation alters her opinions by BC finds herself with opinions that were encoded all the while as conditional probabilities in her original probability distribution. There is a sense in which the probabilities in her most prior probability distribution are incorrigible; the scientist is forever trapped in this distribution, moving from point to point within it but never revising it. She is enslaved to commitments she undertook at the inception of her doxastic career, imprisoned by priors adopted (as Bas van Fraassen[12] says) "at her mother's knee." That these commitments are immune from revision is contrary to contemporary epistemological views and to scientific practice. Creative insights in science often occur precisely when a scientist, without receiving any new evidence or recognizing any logical relationship that had formerly escaped her, breaks out of what had been her most prior probability distribution–say, by deciding to take seriously a hypothesis to which she had been giving an exceedingly small (or zero) most prior probability, or to regard a hypothesis as capable of being confirmed by a class of evidence that she had formerly considered irrelevant to it.[13]

For instance, imagine a physicist of 1905 who is re-reading Albert Einstein's new paper "On a Heuristic Viewpoint concerning the Production and Transformation of Light."[14] Einstein examines the fact that Max Planck's equation governing the black-body spectrum can be derived from the hypothesis that light comes in discrete quantities (quanta) rather than continuous waves, and proposes that this equation's empirical correctness be regarded as evidence for any other prediction that can also be derived from the light-quantum hypothesis. Suppose that on his first reading of this paper, our physicist shared the reaction of nearly all actual physicists in 1905: if the black-body spectrum behaves as if light is quantized, this is sheer coincidence–"nothing more than a curious property of light, without any physical significance."[15] It is not really a live option that light is quantized, considering the overwhelming success of the wave theory of light in accounting for a wide range of optical and electromagnetic phenomena. The light-quantum is just a formal device for deriving the correct black-body equation (as Planck[16] later reported himself to have held in 1905). Accordingly, it would be utterly coincidental if any other phenomenon were as if light is quantized. So the black-body equation's success does not confirm the other equations (governing the photoelectric effect, the photoionization of gases, and photoluminescence) that Einstein uses the light-quantum hypothesis to derive. This, I said, was our physicist's initial reaction. But suppose that on re-reading Einstein's paper, our physicist is more impressed by certain considerations which he noted before, but which (he now believes) he originally failed to accord their proper weight. For example, he is struck more forcibly by the paper's opening thought: that none of the remarkable successes of the classical theory of light (reflection, refraction, diffraction, interference, and so on) concerns light absorption or emission, which are central to each of the phenomena with which Einstein is concerned (the black-body spectrum, photoelectricity, and so on). Although he knew this point before re-reading the paper, our physicist decides that he failed then to give it due regard. He therefore re-evaluates certain of his opinions–for example, raising his pr(light-quantum hypothesis) and setting pr(Einstein's photoelectric-effect equation | Planck's black-body equation) > pr(Einstein's photoelectric-effect equation), whereas he had formerly set these equal. He is not taking into account evidence (or even logical truths) that he did not know or take into account before, nor is he inventing some new hypothesis. Rather, he regards his new probabilities as the *right* ones for him to assign–the ones he *should* have assigned earlier, considering what he knew then. (I shall elaborate this idea of rightness in the next section.)

---

[12] *Laws and Symmetry* (New York: Oxford, 1989) and "Rationality Does Not Require Conditionalization," in E. Ullmann-Margalit, ed., *The Israel Colloquium Series in the History, Philosophy, and Sociology of Science*, Volume V (Dordrecht: Kluwer, forthcoming).

[13] Levi lodges a similar objection against BC, recognizing that rationality leaves room for violations of "confirmational tenacity," as Levi aptly terms the perpetual grip of our most prior probability distribution (*op. cit.*, p. 82). (For more on this, see section II.) See also Howson and Alan Franklin, "Bayesian Conditionalization and Probability Kinematics," *British Journal for the Philosophy of Science*, XLV (1994): 451-66, here p. 459; Charles Chihara, "The Howson-Urbach Proofs of Bayesian Principles," in Eells and B. Skyrms, eds., *Probability and Conditionals* (New York: Cambridge, 1994), pp. 161-78, here pp. 171-73; and especially van Fraassen (*op. cit.*).

[14] *Annalen der Physik* (ser. 4), XVII (1905): 132–48.

Now perhaps it is not initially evident how the problem of incorrigibility is a problem for BC. Our physicist has shifted his opinion, but did not discover any new evidence regarding light quantization, whereas BC purports to cover only shifts involving new evidence. Yet we might just as well have imagined an example where a scientist's creative insight–her breaking out of her most prior probability distribution–occurred together with her discovery of some new evidence, and so her shift of opinion falls within BC's jurisdiction, but violates BC. Indeed, our physicist's shift of opinion with regard to light quantization surely happened to be accompanied by his learning *something*–perhaps that the dog is barking–and so he updated his opinions in violation of BC. Thus, I believe that it is permissible for updates to violate BC.[17]

---

[15] Abraham Pais, *Subtle Is the Lord* (New York: Oxford, 1982), p 337

[16] From a 1931 letter to R.W. Wood, in A. Hermann, ed., *Fruhgeschichte der Quantentheorie*, Claude W Nash, trans. (Cambridge: MIT, 1971), pp. 23-24.

We might amend BC by explicitly restricting its application to cases where we should not break out from our most prior probability distribution. But since $\mathrm{pr}_{\mathrm{new}}(h) = \mathrm{pr}_{\mathrm{old}}(h|e)$ follows from our retaining our most prior probability distribution (in particular, from $\mathrm{pr}_{\mathrm{new}}(h|e) = \mathrm{pr}_{\mathrm{old}}(h|e)$), BC is trivialized by being so amended; it then says merely that we should update by $\mathrm{pr}_{\mathrm{new}}(h) = \mathrm{pr}_{\mathrm{old}}(h|e)$ in cases where we should. Instead, we might amend BC by restricting its application to cases where we have no reason to depart from our most prior probability distribution, and by offering further rules that govern what count as good reasons for so departing.[18] But BC can then account only for those shifts of opinion where the agent should remain within her most prior probability distribution. This apparently threatens all of the familiar Bayesian accounts of various features of confirmation, since these accounts will not apply unless the agent must remain within her most prior probability distribution, and until we have developed some further rules, we cannot explain when this is. I shall argue, however, that although our shifts of opinion need not leave us within our most prior probability distribution, the steps in an argument justifying our current opinion should do so. BC can then be used straightforwardly–without amendment–to explicate '*e* is evidence for *h*', thereby preserving the familiar Bayesian accounts intact.

## II

Here, then, we have four problems that arise when we try to use BC to explicate '*e* is evidence for *h*'. I shall suggest that these problems fail to arise if our explication depends not on BC's governing how we should update our opinions from one moment to the next, in light of new evidence, but instead on BC's governing how we should justify our current opinions by appealing to evidence that we have already accumulated. In showing an audience how her current degree of confidence in *h* constitutes a justified response to the evidence that she already has on hand, a scientist explains the bearing of various pieces of evidence on *h*, but does not change her opinions in the course of doing so.

---

[17] Again, here I concur with van Fraassen, *op. cit.*

[18] Levi (*op. cit.*) allows confirmational commitments to be revised in various ways depending upon the context of inquiry. He offers an account of how we ought to revise our "confirmational commitments" in certain important contexts.

She begins her justificatory argument, I suggest, by appealing to a probability distribution that other scientists should recognize as unbiased in the relevant respect. She is entitled to begin with this distribution as a "free move"–that is, without justifying it.[19] She then justifies her current degree of confidence in *h* by invoking various pieces of evidence that have already been discovered, and by using them successively to "update" the initial, unbiased opinion until all of the relevant evidence has been considered, at which point her current degree of confidence in *h* should be reached. Once again, I put scare-quotes around 'update' because the scientist is not changing her degrees of belief in the course of this exercise. Rather, she is justifying the opinion that she was already holding when she began to demonstrate that she was entitled to hold it. (In the next section, I shall argue that this "updating" must accord with BC.)

For instance, consider how a weather forecaster might justify her current 80% confidence in the hypothesis that it will rain and not snow tomorrow. She begins her justificatory argument with an "unbiased" probability distribution, which (let us suppose) assigns 20% confidence to 'It will rain and not snow tomorrow' (hypothesis $h_1$), 10% confidence to 'It will rain and snow tomorrow' ($h_2$), 5% confidence to 'It will snow and not rain tomorrow' ($h_3$), and so on. In the first step of her argument, she submits her first piece of evidence: that today's barometric pressure was 29 torr. This evidence, she explains, warrants "updating" her opinions in a certain way; among her "intermediate conclusions" at this first step are $\text{pr}(h_1) = .6$, $\text{pr}(h_2) = .10$, and $\text{pr}(h_3) = .01$. Today's barometric pressure, then, is evidence for $h_1$ and against $h_3$, while it counts neither for nor against $h_2$. In the second step of her justificatory argument, she takes into account another piece of the evidence–perhaps today's temperature–she has on hand. She explains the bearing of this evidence on the leading hypotheses; among her intermediate conclusions at this second step are pr($h_1$) = .7, pr($h_2$) = .002, and pr($h_3$) = .001. Today's temperature is evidence for $h_1$ and against $h_2$ and $h_3$. She continues in this fashion until she has submitted all of the evidence she has acquired and deems relevant. For the justificatory argument to succeed, she must at this final step have reached pr($h_1$) = .8, her current opinion.

---

[19] To invoke this distribution's "impartiality" is obviously to issue a giant promissory note. Of course, its impartiality cannot depend on its somehow according all conceivable competing hypotheses equal weight. Later, I say a bit more about impartiality, but to pursue it much further would amount to discussing the problem of the priors, which I shall not do. Let me emphasize that, if my position ultimately requires an account of what makes some probability distribution able to serve as the initial distribution, so does the view that BC governs the revision of opinion in the face of new evidence. I criticized that view as unable, for various reasons, to explicate '*e* is evidence for *h*', but not for presupposing an answer to the problem of the priors; the criticisms I reviewed would still apply even if we had some account of why certain opinions constitute a reasonable place to begin one's doxastic life. My purpose here is to find some way to preserve BC's apparent successes while avoiding the problems I have mentioned, not to resolve the problem of the priors.

Changing your opinion upon ascertaining new evidence is utterly distinct from justifying your current opinion by appealing to evidence that you have already accumulated. The justification of a scientist's current opinion is not necessarily the history of her arrival at that opinion, even if she arrived at that opinion justly. For instance, the scientist may at one time have revised her opinion regarding the hypothesis by considering new evidence that she later decided was false. In the argument that now justifies her degree of confidence in that hypothesis, there is no step in which her opinion is "updated" in light of this evidence, since she now believes this evidence mistaken. Furthermore, her current justification may differ from her earlier justification even if the opinion being justified and the evidence justifying it are unchanged. She may have changed her mind in the meantime regarding whether certain observations count as evidence for or against the hypothesis, or how powerfully they count. Her current justificatory argument takes each successive piece of evidence into account in precisely the manner that she now believes proper.

The distinction between updating your opinions upon receiving new evidence and showing how the evidence on hand justifies your current opinions should not be confused with various other, more familiar distinctions, such as that between the "context of discovery" and the "context of justification." Admittedly, both distinctions have a diachronic side concerned with arriving at something and a synchronic side concerned with justifying something. A crucial aspect of the traditional distinction between the two contexts, however, is that when we discuss some hypothesis in the context of discovery, we make no normative judgments; we describe how (for example) a conversation, a piece of evidence, a dream, or a heuristic procedure led to the formulation of the hypothesis. In contrast, *both* sides of my distinction are normative: how you *ought* to revise your past opinions in response to new evidence versus how your current evidence *warrants* your current opinions. The context of discovery includes various influences, such as dreams, that appear on neither side of my distinction, since they play no justificatory role.[20]

One way to grasp the role in justificatory arguments that (I shall ultimately suggest) BC plays is through the distinctions that Levi carefully draws among the various roles that BC might play in connection with updating opinions. Levi presumes that all rational agents adopt "confirmation commitments" that specify, for each deductively closed set of claims, the opinions that the agent is committed to holding in the event that the claims to whose truth she is committed are exactly the members of that set (*op. cit.*, p. 79). Levi argues that our confirmational commitments at a given moment are constrained by "confirmational conditionalization"–roughly, that we are committed to updating our opinions in accordance with BC (*op. cit.*, pp. 81, 222). Again very roughly: for a consistent expansion of our current full beliefs, formed by adding $e$ to those beliefs and then rendering the result deductively closed, "confirmational conditionalization" requires that we be committed to our $\mathrm{pr}_{\mathrm{old}}(h|e)$ being equal to our $\mathrm{pr}_{\mathrm{new}}(h)$. If our confirmational commitments are constrained by confirmational conditionalization and remain unchanged upon a consistent expansion of our current full beliefs, then our new opinions are dictated by BC. Such a shift of opinion is what Levi calls a "temporal credal conditionalization." But Levi denies "confirmational tenacity" (thereby recognizing what I have called the problem of incorrigibility) (*op. cit.*, p. 82); that is, he denies that a rational agent must always remain faithful to her former confirmational commitments, and so carry out their dictates on the occasion of (say) a consistent expansion of her beliefs in light of an observation. Rather, she may have good reason for revising her confirmational commitments.[21] In that case, her shift of opinion, upon a consistent expansion of her current beliefs, will not generally be a temporal credal conditionalization.[22] Now, rather than consider an agent's shift of opinion from one moment to the next, let us consider her transitions from one step to the next in a justificatory argument. At each step, she has arrived at certain intermediate conclusions along with (let us suppose) certain "confirmational commitments" that specify, for any claim *e* that is consistent with all of the full beliefs already arrived at by that step, the opinions that represent the right way to take *e* into account, should *e* be the further evidence submitted in the next step of the justificatory argument. The analog of "confirmational conditionalization" requires that at each step, the agent's confirmational commitments be governed by BC, in the following sense: if the agent is required to honor her confirmational commitments in moving to the next step in the justificatory argument (the analog of "confirmational tenacity"), and if every transition from one step to the next results from taking account of a further piece of evidence *e* that is consistent with all of the full beliefs arrived at thus far in the argument, then the analog of "confirmational conditionalization" requires that the agent's intermediate conclusion $\mathrm{pr}_{\mathrm{new}}(h)$ at the next step equal her intermediate conclusion $\mathrm{pr}_{\mathrm{old}}(h|e)$ at the current step. This "update" is the analog of a temporal credal conditionalization. In effect, I defend these analogs of confirmational conditionalization and tenacity, and argue that all transitions from one step to the next in a justificatory argument are analogs of temporal credal conditionalizations.

---

[20] Some authors have argued that normative judgments can be made regarding a scientist's means of thinking up hypotheses to entertain. But even if there is a "logic of discovery," BC could play no role in it, since BC operates only on hypotheses that have already been formulated.

[21] See footnote 18.

[22] Levi notes that, even if the agent retains her confirmational commitments, it does not follow that all of her shifts of opinion are temporal credal conditionalizations, since perhaps a legitimate revision may be occasioned by a contraction rather than an expansion of her beliefs, or by new evidence that does not involve learning the truth of any claim. See footnote 10.

In other words, I have suggested (in connection with the problem of incorrigibility) that a scientist is not obliged to revise her opinions, *from one stage in her work to the next*, in accordance with BC. Nevertheless, I shall argue that *at each stage in her work*, BC must govern the steps in the arguments by which her opinions then are justified. In that event, BC would be in a position to account for various features of confirmation without encountering the four problems that I have mentioned.

For instance, the problem of old evidence—as a difficulty for using BC to explicate '*e* is evidence for *h*'—is avoided if this explication applies BC not to updatings, but to the transitions between the steps in justificatory arguments. Although our confidence in *h* was never *updated* in light of *e*'s discovery, since *e* was discovered before *h* was formulated, the argument by which we *now justify* our pr(*h*) appeals to *e*. That argument begins with a probability distribution in which *h* is assigned some nonzero probability and *e* is assigned a probability less than 1 (since this distribution did not take account of any of the evidence to be submitted in the course of the argument). So in the context of a justificatory argument, none of the evidence being presented for (or against) *h* is "old."

Notice that this response to the problem of old evidence is similar to but distinct from the appeal to counterfactual degrees of belief (which I mentioned earlier). A prior probability in a justificatory argument is like a counterfactual degree of belief in that it may be a fictitious state of opinion; perhaps at no moment did our opinions match this prior probability distribution. As in the appeal to counterfactual degrees of belief, the strategy is to argue that there is a reasonable prior that does not construe the old evidence as old. But unlike the counterfactual approach, the prior probability distribution in a justificatory argument is not beholden to the historical facts that determine the confidence that we would have placed in the hypothesis, had it been formulated sometime earlier. These historical facts are irrelevant to our current judgments of whether *e* is evidence for *h*. A given prior probability distribution is relevant to whether *e* is evidence for *h* not because, in the actual world or some possible world, we hold this probability distribution and then raise our confidence in *h* when updating our opinions upon discovering *e*. Rather, it is relevant because this probability distribution figures in the arguments by which we justify our opinion regarding *h*.

It might be objected that I have not removed the problem of old evidence, but merely presupposed that there is a solution to it–that there is a reasonable prior that does not construe the old evidence as already known. But this charge is unfair. Once we shift our attention from updates to justificatory arguments, the temporal element is removed, and so old evidence presents no peculiar problem. The *oldness* of the old evidence presents no special reason for it to be depicted by the initial probability distribution in a justificatory argument as having already been ascertained; a piece of evidence submitted in the course of the justificatory argument, whether or not it was discovered before the given hypothesis was formulated, was perforce found before the justificatory argument was offered. So there is no longer any problem of old evidence. Of course, there remains the question: Whence comes the initial probability distribution in the justificatory argument? But we have not merely exchanged the problem of old evidence for this problem, since this problem was already with us as a version of the problem of the priors.[33]

The second problem I raised–how to use BC to explicate *e*'s serving as evidence for (or against) *h*, if *h* is already believed (or rejected) by the time *e* is discovered–is again not a problem once we stop trying to find an updating to which BC applies, and instead apply BC to the step where *e* is taken into account in a justificatory argument. After we adopt (or reject) *h*, we continue to be committed to arguments that justify our doing so. Suppose that even before discovering *e*, we justly adopted *h*; we already had sufficient evidence to fund an argument justifying belief in *h*. Then *e*'s discovery may nevertheless enable us to give a new argument for believing *h*, one step of which involves taking *e* into account. If that step involves an increase in our confidence in *h*, then it is correct to say (in connection with this argument) that we take *e* as evidence for *h*. In short, there is only one history of the shifts in our opinions upon discovering new evidence, but we may have multiple arguments for our current degree of confidence in *h*. So even if *e*'s discovery resulted in no further shift in our confidence in *h*, nonetheless *e* may count as evidence for (or against) *h*, so long as this is cashed out in terms of *e*'s role in our justificatory arguments.

---

[33] See footnote 19.

The third problem I raised—that BC cannot explain what makes *e* evidence against *h*, if *h* is initially believed and then *e* leads us to reconsider—likewise can be avoided if BC is applied not to *e*'s updating our confidence in *h*, but to *e*'s role in an argument justifying our current degree of confidence in *h*. That is because no claim accepted at one step in a justificatory argument is rejected at a subsequent step. For instance, if a scientist initially accepts a given claim as an observation report but later rejects it, she thenceforth ceases to submit that claim as an observation report anywhere in her justificatory arguments. She does not cite it at one stage in an argument, as evidence relevant to various hypotheses under consideration, and then take it back at a subsequent stage. As Thomas Kuhn[34] emphasizes, the justificatory argument in a scientific paper is "cleaned up"; it does not double-back on itself in the way that a scientist's actual doxastic history might.

Similarly, the problem of incorrigibility is avoided when BC is applied to the steps in a justificatory argument. Our 1905 physicist changed her mind about whether the correctness of Planck's blackbody equation counts as evidence for Einstein's photoelectric-effect equation. In contrast, once having submitted *e* at some step in a justificatory argument as evidence for (or against) *h*, we cannot at some other step in the argument revise our opinion of whether *e* counts as evidence for or against *h*. The argument, I shall suggest, operates within a fixed most prior probability distribution (though we may, of course, later reject the argument). For example, suppose you are justifying your current 75% confidence in *h*. After the first few steps in your argument, you have taken into account a certain portion of your current evidence and reached $\mathrm{pr}(h) = .4$ as your confidence so far. After the next few steps, you have taken into account more of your current evidence and reached $\mathrm{pr}(h) = .8$ as your intermediate conclusion. At this step, you cannot reject your "earlier" degree of confidence as having been the wrong intermediate conclusion to have reached from the evidence under consideration at that "earlier" step. *Since the intermediate conclusions at "earlier" steps in a justificatory argument are the basis for the intermediate conclusions at "later" steps, a "later" step cannot involve the rejection of an "earlier" intermediate conclusion as having been an inappropriate conclusion to have reached from the evidence under consideration at that step.* The opinions appearing as intermediate conclusions at a given step of the justificatory argument must be regarded *throughout the argument* as having been merited by the evidence taken into account at, or prior to, that step.

---

[34] *The Essential Tension* (Chicago: University Press, 1977), p. 327.

I shall now try to spell out this view by using an argument from calibration to show that the steps in a justificatory argument must accord with BC.

## III

A standard way of arguing for BC is by a Dutch Book argument. A *Dutch Book* is a series of bets, each bet deemed fair by the bettor at the time she makes it, such that the bettor will lose money no matter what the outcomes of the events that are the subjects of these bets. In a *synchronic* Dutch Book, the bettor makes all of the bets at once; in a *diachronic* Dutch Book, she makes them at different times. The idea is that, if the bettor's degrees of belief regarding the outcomes of various events are such that the bookmaker can design bets on these outcomes which the bettor will take (that is, that she will deem fair) but which must necessarily result in the bettor's suffering a net loss, then the bettor must have made a logical error in the management of her opinions. It is required that the bookmaker be able to design the Dutch Book even though the bettor and bookmaker know exactly the same facts. For instance, if the bettor uses a rule for updating her opinions upon receipt of evidence, then since she must know the rule she uses, the bookmaker must know that rule, too, and if the bettor does not know the outcomes of the events on which she has bet, then the bookmaker must not know them either. (Obviously, if the bookmaker knows the outcome of the event on which the bettor is betting, and the bettor does not, then it signifies no logical flaw in the bettor's management of her opinions that the bookmaker can construct a bet which the bettor believes fair but from which she cannot profit.) It was shown (by John G. Kemeny, R. Sherman Lehman, and Abner Shimony, with key contributions from Frank Ramsey and de Finetti) that you are vulnerable to a synchronic Dutch Book if and only if your degrees of belief at a given time violate the axioms of the probability calculus, and it was shown (by David Lewis) that you are vulnerable to a diachronic Dutch Book if and only if you update your degrees of belief by using a rule other than BC.

But we cannot use a diachronic Dutch Book argument to show that the steps of a justificatory argument must conform to BC. We might worry that a Dutch Book scenario is not an apt metaphor for an agent's offering a justificatory argument, since an agent does not receive new evidence and revise her opinions in the course of justifying her current opinions. But even setting this worry aside by treating the "updates" in a justificatory argument as like diachronic shifts of opinion, there is a further problem: a Dutch Book argument cannot show that we must update according to BC, since we can always frustrate the bookmaker by using no rule at all to update our opinions. The Dutch Book scenario presumes that, if we do not know in advance how we would revise our opinions under various circumstances, then the bookmaker does not know this either, and so cannot construct a Dutch Book against us. Using no rule is not an empty threat; we saw our 1905 physicist use no rule in changing his mind about whether the success of Planck's black-body equation confirms other equations derivable from the light-quantum hypothesis. Accordingly, van Fraassen[illegible] maintains that, although a Dutch Book argument demonstrates that the only rational *rule* for updating is BC, it permits us rationally to depart from BC so long as we employ no rule at all for updating our opinions. Why, then, is a justificatory argument unsuccessful if its steps depart from BC?

Let us begin with a logically prior question. Suppose we present a justification of our current degree of confidence in a hypothesis, proceeding from an "unbiased" most prior distribution by a sequence of updates. Each update responds to one piece of evidence that we have accumulated, and the sequence of updates continues until all of the evidence on hand that we deem relevant has been considered. Why does this exercise possess the power to justify our current opinion? Why do we succeed in justifying our current opinion by showing that it is guided by the degrees of confidence in the prior step, which are guided by the degrees of confidence in the yet prior step, and so on back to the "unbiased" most prior probabilities? Why does guidance by those intermediate conclusions and initial opinions matter?

It matters, I suggest, because we believe the intermediate conclusions and initial opinions to be "perfectly calibrated." Consider a weather forecaster who has 80% confidence that it will rain tomorrow. Suppose that she bases this opinion on the fact that it rained today, the high temperature today was 60 degrees, and so on. Any other day where it rains on the previous day, the high temperature on the previous day is 60 degrees, and so on–in other words, any other day that is

---

[illegible] *Laws and Symmetry*, p 174; "Rationality Does Not Require Conditionalization."

similar to tomorrow, in all of the respects that grounded the forecaster's opinion regarding rain tomorrow–I shall term *relevantly similar* to tomorrow. Suppose that the weather forecaster knows regarding some other day only that it rained on the previous day, the high temperature on the previous day was 60 degrees, and so on. Then the weather forecaster should have 80% confidence that it will rain on that day, since she has exactly the same reasons in the two cases for believing that it will rain. In assigning 80% confidence to the claim that it will rain tomorrow, the forecaster is perfectly calibrated if and only if it rains on exactly 80% of the days that are relevantly like tomorrow. If it rains on more than 80% of those days, then the forecaster's subjective degree of belief is too low; if it rains on less than 80% of those days, then her opinion is too high. If she is perfectly calibrated, then her opinion is right; it is just what it ought to be, considering her reasons for it (that is, her sense of relevant similarity).[illegible]

In a justificatory argument, we base our intermediate conclusion '$\mathrm{pr}(h) = r$' on certain considerations that we have already submitted by that step. Suppose that we consider some other hypothesis $k$, discussed at that step, to be relevantly similar to $h$, in that we take the considerations on which we base our intermediate conclusion '$\mathrm{pr}(k) = s$' at that step to be relevantly similar to those on which we base '$\mathrm{pr}(h) = r$' at that step. That is, we take ourselves to have exactly the same kinds of reasons for our degree of confidence in $k$ as we do for our degree of confidence in $h$. (Accordingly, $r = s$.) The intermediate conclusions at a given step are "perfectly calibrated" exactly when for each of these conclusions '$\mathrm{pr}(h) = r$', $r$ is the fraction of true hypotheses among the hypotheses that at this step we deem relevantly similar to $h$. I suggest (as a first approximation) that we believe the intermediate conclusions at a given step of a justificatory argument to be perfectly calibrated, since this would explain *why* we believe that we ought to be guided by those conclusions when we move to the next step, taking some further evidence into account. (In a moment, I shall examine just *how* we should be so guided.)

Let me elaborate. Suppose we have no confidence in the calibration of a weather forecaster's predictions. For example, she has 80% confidence that it will rain tomorrow, but our opinions regarding the frac-

---

[illegible] For a similar conception of an opinion's rightness, see van Fraassen, "Calibration: A Frequency Justification for Personal Probability," in R.S. Cohen and L. Laudan, eds., *Physics, Philosophy, and Psychoanalysis* (Boston: Reidel, 1983), pp. 295-319, here p. 301; *Laws and Symmetry*, p. 177. I shall qualify this idea shortly. As van Fraassen emphasizes, this conception of an opinion's rightness does not suggest that our subjective degree of confidence in rain tomorrow is actually to be understood as a claim concerning the objective chance of rain then.

tion of rainy days among the days that she takes to be relevantly like tomorrow are no different from our opinions regarding the fraction of rainy days among the days that she takes to be relevantly like the day after tomorrow, even though she has 80% confidence that it will *not* rain on that day. Then we do not take her forecasts seriously; we give them no weight in guiding our opinion regarding whether it will rain.[27] In a justificatory argument for our current opinion, by contrast, each step's intermediate conclusions are guided by the previous step's; we take those intermediate conclusions seriously when we take some further evidence into account. If we change our minds about the right intermediate conclusions to draw at a given step, then we change that step, adjusting the conclusions there accordingly. We do not retain the step with the intermediate conclusions that (we now judge) fail to accord certain considerations their proper weight, and add to the argument a further step to negate the effect of these conclusions on subsequent steps. To include a step where no additional evidence is submitted, but new intermediate conclusions are drawn to replace the old, would be in effect to start the argument over in the middle by setting aside the intermediate conclusions drawn in the earlier step as having been mistaken there. (I present an example of this in a moment.) The simplest rational reconstruction of why the intermediate conclusions in a justificatory argument are taken seriously throughout the argument, each step's guiding the next step's, is that the agent believes each step's intermediate conclusions to be perfectly calibrated.[28]

In contrast, we *can* rationally change our opinions *from one moment to the next* without ascertaining any new evidence. When we decide that our earlier opinions failed to accord certain considerations their proper weight, we decide that those opinions were not the right ones for us to have held, considering the evidence we then knew. We are taking into account no evidence (or even logical truth) that we formerly failed to take into account; rather, we are changing our minds about *what it takes* for us to take proper account of some consideration that we already knew.

So the key difference between updatings from one moment to the next and "updatings" from one step in a justificatory argument to the next is that in the justificatory case, but not in the diachronic case, we must use the earlier opinions to guide their revision; we deem the later opinions to be right on the basis of deeming the earlier opinions to have been right. If we decide that some intermediate conclusion was not right, then we change that step, giving ourselves a new justificatory argument. But if we change our minds about whether our opinions at an earlier moment accorded proper weight to the evidence then known, we cannot go back and change which opinions we held in the past.

---

[27] For a similar view, see Maurice Fréchet, "Les Définitions Courantes de la Probabilité," *Revue Philosophique*, CXXXVI (1946): 129-69, here p. 144.

[28] Here is a less simple but perhaps more realistic alternative. Rather than require that the agent believe $r$ to be the fraction of true hypotheses among the hypotheses relevantly like $h$, require that $r$ be the agent's *expectation value* for that fraction. That is, require that $r = \Sigma \mathrm{pr}(r_i) r_i$, where $\mathrm{pr}(r_i)$ is the agent's subjective probability that $r_i$ is that fraction. This modification is compatible with the subsequent argument, *mutatis mutandis*.

To see how the later steps of a justificatory argument are supposed to be guided by the earlier steps, so that the earlier steps must be considered throughout the argument to have been right, consider this example. Suppose that a weather forecaster sets out to justify her 80% confidence in rain tomorrow. She begins her argument with an "unbiased" most prior probability distribution. She then considers a piece of evidence: that today's barometric pressure is 29 torr. She assigns tomorrow to a reference class: the set of days each of which is preceded by a day with barometric pressure close to 29 torr. She judges the cases belonging to the same reference class to be "relevantly similar," that is, similar in all of the respects on which she bases her intermediate conclusion at this step regarding whether it will rain tomorrow. Accordingly, in the intermediate conclusions at that step, she assigns the same degree of confidence to rain on each of these days—let us say 50%. She then considers another piece of evidence: whether or not it rained today. She assigns tomorrow to a narrower reference class (taking a different set of cases to be relevantly similar) and accordingly adopts a new intermediate conclusion. Now suppose that, if her second piece of evidence is that it rained today, the weather forecaster is prepared to make her second step's intermediate conclusion 'pr(rain tomorrow) = .2', and if her second piece of evidence is that it did not rain today, the weather forecaster is prepared to make her second step's intermediate conclusion *again* 'pr(rain tomorrow) = .2'. This would be tantamount to her changing her confidence in rain tomorrow from one step to the next without any additional evidence as motivation; plainly, she is not motivated by taking into account that it did rain, as the opposite discovery would have made no difference to her second step's intermediate conclusion, and she is not motivated by taking into account that it either rained yesterday or it did not, as this is a trivial truth and so does not count as new evidence. To go in this way from 50% to 20% confidence in rain tomorrow is to fail to take the first step's intermediate conclusion seriously; she is not being guided by her first step's intermediate conclusions in reaching her second step's. That is not the way a justificatory argument proceeds.

The same flaw occurs even if the two possible intermediate conclusions ('pr(rain tomorrow) = …') in the second step are not identical, if they are both greater than .5 (the first step's intermediate conclusion) or both less than .5. In either case, the agent is not taking the first step's intermediate conclusion seriously. For instance, if both of the second step's possible intermediate conclusions assign less than 50% confidence to rain tomorrow, then she cannot regard it as being right to assign 50% confidence in the first step; she was prepared to lower her pr(rain tomorrow) in the second step no matter what the evidence concerning rain today, so she must regard the first step's 50% assignment as excessively high. This logical flaw is manifested in the fact that the agent cannot with consistency regard her first step's intermediate conclusions as perfectly calibrated and her second step's intermediate conclusions as perfectly calibrated. For if rain occurs in less than 50% of the days following 29-torr rainy days, and occurs in less than 50% of the days following 29-torr nonrainy days, then it cannot be that rain occurs in 50% of the days following 29-torr days.[29] In failing to take her intermediate conclusions seriously, the weather forecaster betrays her belief that they are not calibrated.

Let me elaborate the notion of a "reference class" as it pertains to calibration. For me to believe my intermediate conclusion 'pr($h$) = $r$' to be perfectly calibrated is for me to believe that $r$ is the fraction of truths among the hypotheses that in this step I believe "relevantly similar" to $h$ (and so to which I assign the same degree of confidence as I do to $h$). Thus, my belief in the perfect calibration of my intermediate conclusion is meaningful only if there is a meaningful sense of hypotheses "relevantly like" $h$. Readers may now recall the notorious *problem of the reference class* (or *problem of the single case*) as it afflicts various frequentist analyses of probability: If 'the probability of this coin toss landing heads is 50%' just means that 50% of a large collection of tosses like this one would land heads, what count as tosses *like* this one? Despair over the reference-class problem has often been cited as an argument against any such analysis of probability. But I am not offering an *analysis* or *interpretation* of the probability assignments in justificatory arguments. The opinions that we believe perfectly calibrated are not opinions *about* the fraction of truths among relevantly similar hypotheses; they are subjective probabilities. Moreover, the notion of calibration that I am employing is not afflicted by any problem of the reference class. The hypotheses that I take in this step to be "relevantly similar" to *h* are exactly the hypotheses that are similar to *h* in all of the respects on which I base my intermediate conclusions at this step. In other words, they are the hypotheses that I take myself at this step as having the same sorts of reasons for believing as I do for believing *h*. For the hypothesis of rain tomorrow, it is easy to imagine what it might take at a given step for a hypothesis to be relevantly similar. For another sort of hypothesis, it may be more difficult to characterize the reasons we have for assigning to it the degree of confidence we do.[30] But to make my point, I do not need to offer an account of these reasons. I need presume only that we must have reasons for assigning to hypotheses the degrees of confidence we do (even if we have not explicitly identified these reasons). Unlike the frequentist interpretations of probability I mentioned, talk of perfect calibration does not presuppose some *objective* sense of two hypotheses being "relevantly similar." The characteristics by which *we are guided* in confirming the hypothesis determine its reference class. We take ourselves as being guided by some factors or other (which, once again, we may be able to express in only a rough and ready fashion). I return to this point in the next section.[31]

---

[29] My flawed weather forecaster is intentionally similar to van Fraassen's Piero in "Belief and the Problem of Ulysses and the Sirens," *Philosophical Studies*, LXXVII (1995): 7-37, here p. 15 His example involves a change in opinion from one moment to the next rather than from one step to the next in a justificatory argument. What van Fraassen calls the forecaster's lack of "integrity" I elaborate as her failure to take seriously her first step's intermediate conclusion.

---

[30] Notice that I need not regard every hypothesis to which (at this step) I assign the same degree of confidence as I assign *h* to be "relevantly similar" to *h*. (van Fraassen makes the same point in "Calibration: A Frequency Justification," p. 303; see also the next footnote.) I may consider my reasons for assigning *r* to a hypothesis predicting rain to be quite different from my reasons for assigning *r* to a hypothesis predicting that a heart-transplant patient will survive for five years after his transplant.

[31] van Fraassen offers one way to cash out this notion of the reference classes in terms of which our calibration is assessed. In "Calibration: A Frequency Approach," he says (pp. 302-05) that the reference class for the claim '*x is A*' (perhaps '22 May 1998 will be rainy in Seattle') consists roughly of '*y is A*', '*z is A*', and so on, where for every property *B*, our pr(*x is B*) = pr(*y is B*) = pr(*z is B*), and so on. By restricting the reference class of '*x is A*' to claims attributing *A* to other entities, van Fraassen prevents the reference class for '*x is A*' from consisting of '*x is A*' along with arbitrarily many claims of the form 'The *n*th toss of the die lands ace'. This is important for the following reason:

> [S]uppose that I first state my probability for rain as 1/6 and then you ask me about one thousand tosses of a fair die for the probability of ace and I say 1/6 each time. On the total set of 1001 questions, my personal probability will probably be quite well calibrated, but that reveals nothing about the reasonableness of my initial judgment about rain. To see the problem in acute form, let this first judgment be replaced by two: adding to it also the judgment that the probability of there being no rain equals 1/6 as well. Calibration on the total of 1002 propositions will be quite good, whereas there is something drastically wrong with my probabilities for the first two.
>
> So the possibility of ever better calibration which we require, must be on extensions of the initial set of propositions which are in a relevant sense *like* the original ones (pp. 304-05).

van Fraassen's thought seems to be that a rational agent must have the same reasons for her opinion regarding *y* as for her opinion regarding '*x is A*' if *y* attributes

I have suggested that an argument deriving our current degree of confidence in a hypothesis from an "unbiased" most prior distribution, by a sequence of updates that eventually takes all of our evidence into account, possesses the power to justify our current opinion because we are entitled to believe that this argument's intermediate conclusions are perfectly calibrated. We are entitled—without having to submit any argument—to believe the most prior distribution to be perfectly calibrated; that this is a free move is part of what it is for the most prior distribution

---

*A* to some other entity *y* and if she has the same confidence in '*y* is *A*' as in '*x* is *A*', the same confidence in '*y* is *B*' as in '*x* is *B*', and so on van Fraassen refuses to cash out the reference class by appealing directly to our having *the same reasons* for our confidence in *y* as for our confidence in '*x* is *A*'; he is willing to appeal only to features of our probability distribution.

But this self-imposed limitation leads van Fraassen to resort to the heavy-handed measure of requiring the reference class for '*x* is *A*' to consist exclusively of claims attributing *A* to other entities. It is not at all clear to me how this analysis applies when the claim in question is not of the monadic form '*x* is *A*'—for example, when the claim is quantified. How do we assess whether a nineteenth-century physicist's opinion of Newton's gravitational-force law is calibrated?

I am not as squeamish as van Fraassen is about carving out *h*'s reference class by appealing directly to our reasons for holding our opinion regarding *h*. I have no way to read these reasons off of our probability distribution Analyzing what it means to say that our opinion regarding *h* is *based* on a given consideration is a classic epistemological problem. See, for instance, Keith Lehrer, *Knowledge* (New York: Oxford, 1974), and George Pappas, "Basing Relations," in Pappas, ed., *Justification and Knowledge* (Boston: Reidel, 1979), pp. 51-63. I assume nothing about what our reasons are, only that we must have reasons. (See section IV.) They presumably relate to our probability distribution in something like the following way. Our reasons for assigning a certain degree of confidence $r$ to $h$ consist of our belief that $h$ possesses various properties $\alpha, \beta, \ldots$, where these properties could logically possibly hold of any number of (logically inequivalent) claims. Let $\Gamma$ be a dummy variable standing for a claim. Then consider our $\mathrm{pr}(\Gamma \mid \Gamma$ possesses $\alpha, \beta, \ldots)$. In other words, for an unidentified claim, this is our degree of confidence in it, given various facts about it. So this might be $\mathrm{pr}(\Gamma \mid \Gamma$ says that it will rain on a certain day, this day follows a day where the high temperature was 80 degrees and the barometric pressure was 30 torr,...). Our belief that $h$ possesses $\alpha$, $\beta$, can be our reason for assigning $r$ to $h$ only if our $\mathrm{pr}(\Gamma \mid \Gamma$ possesses $\alpha, \beta, \ldots) = r$, and for any further claim $g$ that we believe, our $\mathrm{pr}(\Gamma \mid g, \Gamma$ possesses $\alpha, \beta, \ldots) = r$. (For some general remarks in a similar spirit, see Wesley Salmon's discussion of the kinds of properties the possession of which by a hypothesis can serve as a basis for a plausibility judgment—in *The Foundations of Scientific Inference* (Pittsburgh: University Press, 1967), pp. 124-26.)

The above constraint on *h*'s reference class ensures, for instance, that for any $k$ in that class, $\mathrm{pr}(k) = \mathrm{pr}(h)$. This is also a feature of van Fraassen's account. But this constraint does not preclude the reference class of '22 May 1998 will be rainy in Seattle' from consisting of arbitrarily many claims of the form 'The *n*th toss of the die lands ace'. I have presented no *logical* bar to someone's having the following reason for assigning a degree of confidence of 1/6 to *j*: either *j* concerns rain in Seattle on 22 May 1998 and the high temperature on the previous day was 80 degrees,..., *or j* concerns a fair die landing ace. *We* presumably have different reasons for our opinion regarding the rain claim, however, than for our confidence in the die claim. (To say just what makes this so, we would need to identify the "basing relation.")

to be "unbiased."[*] Each step of the justificatory argument takes account of another piece of evidence by revising the probability distribution from the prior step so that it is logically possible for the prior step's intermediate conclusions and the new step's intermediate conclusions both to be perfectly calibrated. The agent sees herself as in each step taking account of some additional piece of evidence so as to preserve in her new intermediate conclusions the perfect calibration that began with her most prior probability distribution. She justifies her current opinion by showing herself entitled to take her current opinion to be perfectly calibrated—by showing how her current opinion results from calibration-preserving updates going back ultimately to her most prior opinions, whose calibration she is entitled to take as a working assumption.[*] I think that this picture explains why such an argument possesses the power to justify our current opinion regarding a hypothesis. For if we have shown ourselves entitled to believe a probability distribution to be perfectly calibrated, then (I argued) we have shown ourselves entitled to believe those degrees of confidence to be *right*, and so we have shown ourselves entitled to hold those degrees of confidence. When our 1905 physicist decided that he had previously failed to assign certain hypotheses the degrees of confidence they had merited considering the success of Planck's black-body equation, he was deciding that those earlier opinions had not been perfectly calibrated.

Note that I have argued neither that our current opinions nor that the intermediate conclusions in a justificatory argument must *be* perfectly calibrated in order for them to be justified. I have argued only that we must *justly believe* our current opinions to be perfectly calibrated in order for us to be justified in holding them, and that we justify so believing by showing how those opinions result from calibration-preserving updates, ultimately of a prior distribution that we are automatically entitled to take as calibrated. For us to be justified in believing certain opinions to be calibrated, it is not necessary that they actually be calibrated, but (given logical omniscience) their calibration must be logically possible.

Perfect calibration is nearly but not quite the sense of rightness appropriate for the intermediate conclusions at a given step of a justificatory argument. Suppose that at a step where 'pr($h$) = .8' is an intermediate conclusion, we discuss only one other hypothesis that we deem to be relevantly similar to $h$. It is logically impossible for exactly 80% of these two hypotheses to be true (since 80% of 2 is not an integer). So our intermediate conclusions at this step cannot be (justly believed to be) perfectly calibrated. Clearly, this is not from any shortcoming in those opinions; it is merely a technicality. To elaborate the "rightness" of these intermediate conclusions, we need what van Fraassen[34] calls "potential perfect calibration." Consider again the weather forecaster. Each day she assigns the next day to a certain reference class based on the current barometric pressure, temperature, and so forth. She judges the cases belonging to the same reference class to be "relevantly similar," that is, similar in the respects on which she bases her opinion regarding whether it will rain the next day. Accordingly, she assigns the same likelihood *r* of rain the next day to each of the co-classified cases. Now take the actual cases (days) belonging to a given reference class, some fraction *f* of which are rainy days. Consider what would happen to *f* were hypothetical further cases, relevantly like the actual cases, to be added. Consider what the limiting relative frequency would then be of rainy days among days where the forecaster's meteorological data fall within the relevant range. The forecaster's degrees of confidence are "potentially perfectly calibrated" exactly when for each reference class, *f* would approach arbitrarily close to *r* were the number of cases in the reference class to increase without bound. In an analogous way, we can specify what it is for the intermediate conclusions at a given step of a justificatory argument to be "potentially perfectly calibrated." They are potentially perfectly calibrated if and only if for any hypothesis *h* and number *r*, if '$\mathrm{pr}(h) = r$' is an intermediate conclusion at that step, then among hypotheses deemed at that step relevantly like *h* and accorded a degree of confidence (also *r*) at that step, the fraction of truths would approach arbitrarily close to *r* were there sufficiently many such hypotheses.[35]

---

[*] For any opinions consistent with our evidence, there are surely *some* prior probabilities that would yield these opinions via Bayesian updating on our evidence. But there is no guarantee that these prior probabilities would be "unbiased"

[*] In my examples, the agent shows this by presenting a *sequence* of updates, though she could in principle do so by a *single* update on all of her evidence

As van Fraassen[36] explains, it is logically possible for one's opinions at a given time to be potentially perfectly calibrated only if those opinions satisfy the probability calculus. To glimpse the idea behind his argument, notice that, if we assign greater than 100% confidence to *h*, then it is logically impossible, no matter how many hypotheses we are asked about that we deem relevantly like $h$, for more than 100% of these hypotheses to be true. Likewise, if we assign 60% confidence to $h$ and 60% confidence to $\neg h$, then again, it is logically impossible, no matter how many hypotheses we are asked about that we deem relevantly like $h$ (and so whose negations we deem relevantly like $\neg h$), for the fraction of truths among the hypotheses like $h$ to approach 60% and the fraction of truths among the hypotheses like $\neg h$ to approach 60%, since that outcome would require that certain hypotheses and their negations both be true.[37]

---

[34] "Calibration: A Frequency Justification," pp. 304-05; "Belief and the Will," this JOURNAL, LXXXI, 5 (May 1984): 235-56, here p. 245.

[35] van Fraassen's concept of the limiting relative frequency of truths among relevantly similar hypotheses raises a host of familiar worries (such as how hypotheses are to be individuated, how the counterfactual 'Had there been more relevantly similar cases' is to be understood, whether the limiting relative frequency depends on the order in which the relevantly similar cases are considered, and so on)—familiar since analogous worries arise for certain frequentist views of objective probability. I shall not investigate here whether these worries apply to the use I wish to make of potential perfect calibration, or how they might be addressed.

[36] "Calibration: A Frequency Approach"; "Belief and the Will," pp. 245-46 See also Shimony, "An Adamite Derivation of the Calculus of Probability," in J.H. Fetzer, ed., *Probability and Causality* (Dordrecht: Reidel, 1988), pp. 151-61.

I shall apply van Fraassen's argument to our opinions at a given step in a justificatory argument rather than to our opinions at a given moment. The argument then shows that it is logically possible for the intermediate conclusions at a given step of a justificatory argument to be potentially perfectly calibrated only if they satisfy the probability calculus. This constitutes a synchronic calibration argument. I shall now give a diachronic argument–again, not in the sense of applying to a shift in opinion from one moment to the next, but to a shift in opinion from one step in a justificatory argument to the next, as an additional piece of evidence is taken into account. Suppose $\mathrm{pr}(\cdot)$ is our probability distribution at a given step in a justificatory argument, and suppose $\mathrm{pr}'(\cdot)$ is our probability distribution at the next step in that argument, where $e$ is the additional evidence that is brought to bear. Suppose $\mathrm{pr}(\cdot)$ is potentially perfectly calibrated. I now show that it is logically impossible for $\mathrm{pr}'(\cdot)$ and $\mathrm{pr}(\ )$ both to be potentially perfectly calibrated unless $\mathrm{pr}'(\cdot)$ results from updating $\mathrm{pr}(\cdot)$ according to BC.

---

[37] van Fraassen guarantees that the negation of a claim in $h$'s reference class is in $\neg h$'s reference class: the reference class for '$x$ is $\neg A$' consists of claims of the form '...is $\neg A$', and if $y$ is such that '$y$ is $A$' is in the reference class for '$x$ is $A$', then for every property $B$, our $\mathrm{pr}(x \text{ is } B) = \mathrm{pr}(y \text{ is } B)$, and so the reference class for '$x$ is $\neg A$' includes '$y$ is $\neg A$'. Does the scheme I sketched in footnote 31 require this sort of coordination between the reference classes for $h$ and $j$ where $j = \neg h$? It presumably would, once some account was added of what it means for an opinion to be "based" on a given consideration. But this coordination would be required only when we know all of the relevant logical truths, since only then could our reasons for our confidence in $j$ be our reasons for our confidence in $h$ along with the fact that $j = \neg h$. (In other words, if '$\Gamma$ possesses $\alpha$' distinguishes $h$'s reference class, then '$\Gamma$ possesses $\beta$' can distinguish $j$'s, where $\beta$ is the property of being logically equivalent to the negation of a claim possessing $\alpha$.) I shall shortly discuss this point further in the main text.

In the argument that the logical possibility of potential perfect calibration requires coherence, logical omniscience is presumed (in order for various claims' reference classes to be "coordinated" as above). This seems reasonable, since in the absence of logical omniscience, it is unclear that coherence is required. I assume logical omniscience throughout. But it is interesting to note that on my conception of the reference classes, an agent could be calibrated even if she is ignorant of some relevant logical truths. She might, for instance, assign various number-theoretic hypotheses to various reference classes, depending on the reasons behind the degrees of confidence in $\langle 0,1 \rangle$ that she currently assigns them. There is room here to treat the confirmation of logical truths.

Suppose that $\mathrm{pr}(\cdot)$ is potentially perfectly calibrated, so $\mathrm{pr}(h)$ equals the limiting relative frequency of truths among the hypotheses like *h* in the respects relevant to us at the given step. Roughly put, $\mathrm{pr}(h)$ is the value approached by the fraction of truths in a large set *S* of hypotheses each of which is like *h* in the respects relevant to us at this step. For $\mathrm{pr}'(\cdot)$ to be potentially perfectly calibrated, $\mathrm{pr}'(h)$ must equal the value approached by the fraction of truths in a large set of hypotheses each of which is like *h* in the respects relevant to us at the next step in the justificatory argument. But these respects are exactly the respects that were relevant to us at the previous step plus which *e* holds; the hypotheses like *h* in the respects relevant to us at step $(n+1)$ are exactly the hypotheses for which evidence like *e* holds and which are like *h* in the respects relevant to us at step $(n)$. So for $\mathrm{pr}'(\cdot)$ to be calibrated, $\mathrm{pr}'(h)$ must equal the value approached by the fraction of truths among those hypotheses in *S* for which evidence like *e* holds. Since $\mathrm{pr}(\cdot)$ is potentially perfectly calibrated, $\mathrm{pr}(h \mathbin{\&} e)$ is the fraction, among hypotheses in *S*, of truths for which evidence like *e* holds, while $\mathrm{pr}(e)$ is the fraction, among hypotheses in *S*, of hypotheses for which evidence like *e* holds. That is,

$$\mathrm{pr}'(h) = \frac{\substack{\text{\# of hypotheses in } S \text{ that are true} \\ \text{and for which there is evidence like } e}}{\text{\# of hypotheses in } S \text{ for which there is evidence like } e}$$

$$\mathrm{pr}(e) = \frac{\text{\# of hypotheses in } S \text{ for which there is evidence like } e}{\text{\# of hypotheses in } S}$$

$$\mathrm{pr}(h \mathbin{\&} e) = \frac{\substack{\text{\# of hypotheses in } S \text{ that are true} \\ \text{and for which there is evidence like } e}}{\text{\# of hypotheses in } S}$$

in the limit of an arbitrarily large *S*. Hence, $\mathrm{pr}(h \mathbin{\&} e)/\mathrm{pr}(e) = \mathrm{pr}'(h)$. But by definition, $\mathrm{pr}(h \mathbin{\&} e)/\mathrm{pr}(e) = \mathrm{pr}(h|e)$. Therefore, $\mathrm{pr}'(h) = \mathrm{pr}(h|e)$. Thus it is shown that a justificatory argument must proceed by BC.

This reasoning nearly convinces me. The main source of my hesitation is that this argument depends on *e*'s reference class "meshing" with *h*'s. That is, since $\mathrm{pr}(\cdot)$ is potentially perfectly calibrated, $\mathrm{pr}(e)$ is the limiting relative frequency of truths among claims relevantly like *e*. Is part of what makes a claim relevantly like *e* that there is a corresponding hypothesis relevantly like *h*? The above argument presumes so, since it takes $\mathrm{pr}(e)$ to be the fraction, among hypotheses in *S*, of hypotheses for which there is evidence like *e*. But perhaps this presupposition is implausible.

Is some claim made relevantly like *e* partly by its relation to a hypothesis like *h* in the same way as two days are rendered relevantly similar, for the weather forecaster, partly by their each following days whose barometric readings are relevantly alike? Could a claim be relevantly like *e* but correspond to no hypothesis like *h*, and even if it did correspond to such a hypothesis, could this fact play no part in fixing its reference class?

But perhaps this "meshing" of *h*'s and *e*'s reference classes is exactly right. In the above argument, *h*'s reference class in the step where *e* is taken into account seems clearly to consist of those hypotheses in *S* for which evidence like *e* holds. Just as the previous day's barometric reading helps to fix the reference class to which a given day belongs, so in the later step of the justificatory argument, *h*'s being associated with evidence like *e* helps to fix *h*'s reference class. As *e* is bound up in *h*'s reference class, why should *h* not be bound up in *e*'s? Just as the weather forecaster determines the likelihood of rain on a given day by assigning that day to a reference class of relevantly similar days that she has experienced, so the scientist might determine the likelihood of making observation *e* by noting that *e* is strongly suggested by hypothesis *h* and thereby assigning *e* to a reference class of claims that are all strongly suggested by hypotheses like *h*. Notice that the same coordination of reference classes was required by the argument given earlier that, if we assign 60% confidence to *h* and 60% confidence to $\neg h$, then it is not logically possible for us to be potentially perfectly calibrated. That argument required that the hypotheses relevantly similar to $\neg h$ be the negations of the hypotheses relevantly similar to *h*. This seems correct.

If the reference classes must mesh in this way, then I have shown that the steps in a justificatory argument must conform to BC. In this role, I have argued, BC can be plugged into the familiar Bayesian proposals for explicating various features of confirmation in science. Notice, again, that no analogous argument can be made that scientists must accord with BC in updating their opinions from one moment to the next, upon receiving new evidence. At any moment, I have argued, a scientist must believe her *current* opinions (and the intermediate conclusions in the arguments by which she justifies them) to be potentially perfectly calibrated, but she may believe that her opinions *at other moments* fail(ed) to be potentially perfectly calibrated. She may change her mind–at one time believing, at another moment denying that certain opinions match the objective frequencies.[*]

---

[*] Since she need not believe her past opinions to have been calibrated, the connection I have demonstrated between calibration and BC does not show that, upon receiving new evidence, she must diachronically update her opinions in accordance with BC. It might be argued that an agent must not only *believe* her opinions to be calibrated, but for those opinions genuinely to be justified, they must *be* calibrated. It would follow that

## IV

As I noted, a diachronic Dutch Book argument purports to show that BC is the only acceptable *rule* for updating opinion. It thus leaves room for van Fraassen's position: that we can violate BC without vulnerability to a Dutch Book if we use no rule for updating our opinions. But the above calibration argument leaves no room for this position. It shows that the intermediate conclusions in a justificatory argument can logically possibly all be potentially perfectly calibrated *only if these conclusions succeed each other according to BC.* No reference is made to whether some rule is consulted in proceeding from one step to the next.

It might be objected that my argument *does* presuppose that some rule is involved—namely, a rule fixing which hypotheses are (by the agent's own lights) relevantly similar (at a given step) to *h*. If the agent endorses no such rule, that is, recognizes no sense of hypotheses "relevantly similar" to *h*, then (the objection runs) there is no possibility of her reference classes being such that her opinions cannot logically possibly be potentially perfectly calibrated—because she has no reference classes.

But my suggestion was *not* that, *if* the agent recognizes a sense of "relevant similarity" to *h*, then she must believe that her intermediate conclusion regarding *h* could logically possibly be potentially perfectly calibrated, and so the transitions between the steps of a justificatory argument must accord with BC. Rather, I suggested that a rational agent *must* recognize a sense of "relevant similarity" to *h*. That is because she must believe that her opinion of *h* is potentially perfectly calibrated, since this is to believe it to be *right* (considering the evidence brought to bear by that step). She must believe it to be right since she believes that it should guide her in taking account of some further evidence when moving to the next step in the justificatory argument.

It is worth noting another difference between the role in the Dutch Book argument played by a rule for updating and the role in my ar-

---

when an agent holds justified prior opinions and then, upon acquiring new evidence, adopts justified new opinions, this updating must conform to BC. But my argument that an agent must believe her opinions at a given step of a justification to be calibrated was, in brief, that otherwise she would have no reason to base her opinions at the next step on the opinions at the previous step. This argument does not somehow generalize to show that an agent's opinions must really be calibrated. The opinions at one step of a justificatory argument are based on the opinions at the previous step, but the same does not hold diachronically; an agent need not be guided by her old opinions in forming new ones That is why I am willing to require that an agent *believe* her opinions to be calibrated, but not to require that agent's opinions really be calibrated.

Whether justified opinions must *be* calibrated turns on issues concerning epistemological externalism that are beyond this paper's scope. Of course, any requirement that justified opinions be calibrated would remain crucially subjective, considering that the reference classes depend on which hypotheses *the agent considers* relevantly similar.

gument played by a rule determining which hypotheses are relevantly similar (at a given step) to *h*. The Dutch Book defense of BC requires that there be some particular rule for updating that the agent has explicitly adopted (so the bookmaker also knows in advance how the agent would update, and can design the book accordingly). In contrast, my argument does not presume the agent to have explicitly committed herself to some particular rule for determining which hypotheses are relevantly similar (at a given step) to *h*. Furthermore, the Dutch Book argument presumes that the agent has a disposition to update her opinions in a certain way if she receives certain evidence (and that the agent knows what her disposition is, since the bookmaker must know). In contrast, the calibration argument does not presume that the agent currently has some disposition to believe, under certain circumstances, that a given hypothesis is relevantly similar to *h*. The agent may have no such disposition because she does not recognize much about the sense of relevant similarity upon which she is relying. That is, she may not have identified precisely what her reasons are for her opinion (at a given step) regarding *h*, so she may have no disposition to believe that her reasons in another case, for the opinion she would hold there regarding some hypothesis, would be the same as her reasons (at this step) for her opinion regarding *h*.[39] The calibration argument presumes only that the agent *has* such reasons—that her opinions are grounded in some sense of "relevant similarity." This suffices, I have argued, to require that her justificatory arguments conform to BC.

van Fraassen defends a principle he calls *reflection*: that at any moment $t$, $x$ is your degree of confidence that $h$ is true in the event that at $t$ or some time thereafter, $x$ is your degree of confidence in $h$. (That is, reflection says that for $t \leq t'$, $\mathrm{pr}_t\,(h|\mathrm{pr}_{t'}(h)=x) = x$.) Roughly speaking, reflection expresses an agent's belief that now and in the future, her opinions are calibrated.[40] van Fraassen[41] has shown that an agent who updates her opinions from one moment to the next according to BC must conform to reflection; Phillip Dawid[■] has shown likewise that a coherent Bayesian forecaster must assign probability one to the event that she is and will be potentially perfectly calibrated. van Fraassen and Dawid then ask how such confidence in one's own current and future calibration could be warranted.

---

[39] To *have* reasons, it is not necessary that you be in a position to articulate them, or even to say whether your reasons in one case are the same as your reasons in another. To do that may require careful philosophical thought. (For more on what it takes for an opinion to be *based* on a given consideration, see Lehrer (*op. cit.*) and Pappas (*op. cit.*) and the references therein.) The agent might acquire a disposition to believe that (at a given step) she has the same reasons for her opinions regarding two hypotheses. Such a disposition might be groomed by asking the agent to address questions like 'Why do you have a different degree of confidence in Brown's five-year survival after a heart transplant than Smith's, considering that they are similar in all of these respects...?'

[40] For a careful discussion of what reflection requires, see Mitchell Green and Christopher Hitchcock, "Reflections on Reflection," *Synthese*, XCVIII (1994): 297-324.

[41] "Belief and the Will."

What I showed in the previous section is not van Fraassen's or Dawid's result, but more like their converse. I showed that an agent who is potentially perfectly calibrated before and after an update must have updated in accordance with BC. This result is not quite the converse of van Fraassen's or Dawid's because it concerns whether an agent *is* potentially perfectly calibrated, not whether she *believes* herself so to be. Of course, I have also argued that an agent must believe in the potential perfect calibration of the intermediate conclusions at each step of her justificatory arguments. In particular, then, she must believe that the opinions at the final step of her justificatory argument (that is, her *current* degrees of belief) are potentially perfectly calibrated. But I have not argued that she must believe herself to be calibrated at each *future* moment. On the contrary, such a requirement seems just as counterintuitive as requiring that she believe herself to have been calibrated at each past moment.[°] I have emphasized that an agent who at one time believes certain opinions to be merited by certain evidence may later change her mind. Although BC applied to diachronic updatings entails reflection, I avoid requiring that an agent believe in the calibration of her future selves, since I fail to require an agent to conform to BC in updating her opinion from one moment to the next, as new evidence is received. Instead, I regard BC as governing the steps of her justificatory arguments.

The philosophical challenges here lie not so much in giving the formal argument, at the close of the previous section, as in understanding just what that argument shows. Hence, the bulk of this paper was devoted to arguing (i) that an agent must believe the intermediate conclusions in her justificatory arguments to be potentially perfectly calibrated, and (ii) that BC could avoid certain important difficulties if it was applied within justificatory arguments rather than diachronically. I hope to have suggested a new way that BC might account for various features of theory confirmation in science.

MARK LANGE

University of Washington

---

[■] "The Well-Calibrated Bayesian" (with discussion), *Journal of the American Statistical Association*, LXXVII (1982): 605-13.

[°] van Fraassen addresses objections to reflection in his "Belief and the Problem of Ulysses and the Sirens."

THE JOURNAL OF PHILOSOPHY
VOLUME XCVI, NO. 7, JULY 1999

CALCUTTA UNIVERSITY LIBRARY
PUR
CAMPUS
3-1-2000

# PRESENTISM AND ONTOLOGICAL COMMITMENT*

Presentism is the doctrine that only the present is real. Since ordinary talk and thought are full of quantification over non-present objects, presentists are in a familiar predicament: in their unreflective moments, they apparently commit themselves to far more than their ontological scruples allow.

A familiar response is to begin a project of paraphrase. Truths appearing to quantify over problematic entities are shown, on analysis, not to involve quantification over those entities after all. But I think that we might be better off abandoning paraphrase altogether. I suggest a project of discovering *underlying truths* rather than paraphrases. I shall explore this strategy as applied to defending presentism, but my hope is that lovers of desert landscapes everywhere will herein find words of comfort.[1]

---

* I would like to thank John G. Bennett, David Braun, Rich Feldman, Tamar Szabó Gendler, John Hawthorne, Europa Malynicz, and Dean Zimmerman for their help with this paper. I would also like to thank Earl Conee; much of what I say here about ontological commitment has been influenced by hearing his thoughts on the topic. Finally, I would like to acknowledge a special debt to Ned Markosian: this paper began as a commentary on his presentation to the Philosophy of Time Society (Los Angeles, 1998).

[1] In fact, the general project has more importance to me than the special case, since I do not myself endorse presentism In particular, I make no effort to defend presentism from the objections that it is inconsistent with (i) contemporary physical geometry (Hilary Putnam claims that presentism is incompatible with special relativity—"Time and Physical Geometry," this JOURNAL, LXIV, 8 (April 27, 1967): 240-47; and see also page 336 below); or (ii) the claim that truth is supervenient on being; see David Lewis, review of D. M. Armstrong's *A Combinatorial Theory of Possibility*, *Australasian Journal of Philosophy*, LXX (1992): 218-19.

Desert lovers should also consult Joseph Melia, "On What There's Not," *Analysis*, LV (1995): 223-29; and Stephen Yablo, "A Paradox of Existence," in Anthony Everett and Thomas Hofweber, eds., *Empty Names, Fiction, and the Puzzles of Non-Existence* (Stanford: CSLI, forthcoming); many of their conclusions mesh well with the present approach. In particular, there is much in common between Yablo's approach and my own, of which I became aware only after writing this paper; I cannot undertake a comparison in the present paper.

0022-362X/99/9607/325–47

## I PRESENTISM

A presentist thinks that everything is present; more generally, that, necessarily, it is always true that everything is (then) present.[1]

*Presentism* is the temporal analogue of the modal doctrine of *actualism*, according to which everything is actual. The opposite view in the philosophy of modality is *possibilism*, according to which nonactual things exist; its temporal analogue is *eternalism*, according to which there are such things as merely past and merely future entities.

Where possibilists and eternalists speak with quantification, actualists and presentists must make do with irreducible sentence operators. The operators are modal operators for the actualist–⌜NECESSARILY($\phi$)⌝ and⌜POSSIBLY($\phi$)⌝–and tense operators for the presentist:⌜WAS⌝($\phi$) and⌜WILL⌝($\phi$), among others. Whereas an eternalist can say 'There exists, located in the past, a dinosaur with a 50 foot long tail', a presentist must say 'WAS(there exists a dinosaur with a 50 foot long tail)'. The truth of this sentence is consistent with presentism because the existential quantification occurs within the scope of the tense operator, and thus does not carry a commitment to the existence of a dinosaur, just as 'POSSIBLY(there exists a unicorn)' is taken by the actualist not to imply the existence of a nonactual unicorn.

Given the presentist's acceptance of the tenses, some have wondered whether the dispute over presentism is merely verbal. The presentist says while the eternalist denies that everything is present. But is it clear that each means the same thing by 'everything'? The eternalist might view the presentist's quantifiers as being quantifiers restricted to present entities; likewise, it might be argued, when the eternalist says 'There is a dinosaur', this may be translated into the presentist's language as 'There is, was, or will be a dinosaur'. But the translation scheme implicit in this argument is in general not truth preserving. For example, the eternalist will allow that there exists a set containing a dinosaur and a computer; but the proposed presentist's translation of this assertion is the disjunction:

> (There is a set containing a dinosaur and a computer) or WAS(there is a set containing a dinosaur and a computer) or WILL(there is a set containing a dinosaur and a computer).

---

[1] On presentism, see Arthur N. Prior, "Changes in Events and Changes in Things" and "Quasi-propositions and Quasi-individuals," each in his *Papers on Time and Tense* (New York: Oxford, 1968); John Bigelow, "Presentism and Properties," in James E. Tomberlin, ed., *Philosophical Perspectives, Volume X: Metaphysics* (Cambridge: Blackwell, 1996), pp. 35-52; and Ned Markosian's forthcoming "A Defense of Presentism," in Aleksandar Jokic and Quentin Smith, ed., *Time, Tense, and Reference* (New York: Oxford, forthcoming). For more references, see Bigelow's bibliography and Markosian's note 3.

which is presumably false since at no time does there exist *both* a dinosaur and a computer.[3]

Moreover, if the dispute between presentists and eternalists is merely verbal, then the same ought to be true for the dispute between actualists and possibilists; but, surely, this is not the case.[4] The dispute between the presentist and eternalist, then, is genuine. Each uses the unrestricted quantifier 'everything' in the same way, as applying to absolutely everything; one thinks this includes merely past and future objects, the other does not.

## II PROBLEMS FOR PRESENTISM

I turn now to certain problems facing a presentist, which arise from the fact that we often appear to quantify over merely past things:

(D) Dinosaurs are animals that once walked the earth.

The case of (D) is quite unproblematic; a presentist can provide a translation of (D) that eliminates the apparent quantification over dinosaurs:

($D_P$) WAS (there are some animals that are dinosaurs and walk the earth).

It might be plausible for a presentist to argue that ($D_P$) is what we actually mean when we utter (D), or at least that a charitable semantics would associate the proposition expressed by ($D_P$) with (D). The apparent quantification in (D) over past dinosaurs could be attributed to the fact that ordinary speakers are often careless about the order of quantifiers and sentential operators, particularly when the difference only matters if an esoteric metaphysical doctrine like presentism is true. A large class of sentences about the past and future can be handled in this way.

But some apparently cannot be. (D) is a special case in a couple of ways. One is that it is purely qualitative; there are no proper names, demonstratives, or indexicals referring to particular entities. Another is that (D) talks about the past, as it were, one time at a time; there are no ascriptions of cross-time relations. Departing from either of these features of (D) raises problems for the presentist.

Take, for example:

(L) Lincoln was tall.

The only way to translate (L) into a presentist truth would seem to be to replace 'Lincoln' with some description. Arthur N. Prior takes this route (*op. cit.*, pp. 11-14). (L) might thereby be paraphrased as:

---

[3] I am assuming that the presentist assumes that it is always the case that sets exist only if all their members do.

[4] Yet another reason to think the dispute is genuine is the apparent conflict between presentism and certain empirical theories—see footnote 1.

($L_P$) WAS (there is someone who is called 'Lincoln', who is honest,... and who is tall).

For a translation of (L) to be true, it must express a true proposition. But the proposition cannot be a singular proposition containing Lincoln as a constituent, since Lincoln does not exist. The only remaining possibility seems to be to say that the translation of (L) expresses a purely qualitative proposition, containing just qualitative properties and relations as constituents. ($L_P$) expresses such a proposition. But now the problem is that Saul Kripke and other antidescriptivists have argued powerfully that names do not abbreviate descriptions.[5] ($L_P$) is therefore not synonymous with (L).

Cross-time relations present a quite different problem, which arises from a limitation on the sorts of fact that can be expressed in the tensed language used by the presentist. Heuristically, one may think of the sentence $\ulcorner\mathrm{WAS}\urcorner(\phi)$ as being true with respect to a time, *t*, if and only if sentence $\phi$ is true with respect to some time before *t*. And, still heuristically, a sentence $\phi$ (that lacks tense operators) is true with respect to some time if and only if it is true when the quantifiers in $\phi$ range over objects that exist at that time, and atomic formulas are evaluated with respect to that time. Of course, the presentist cannot accept this explanation of the tense operator, as the explanation quantifies over nonpresent times and objects; for the presentist, the tenses are primitive.[6] Nevertheless, the heuristic device gives the idea, much in the same way that the idea of possible worlds clarifies modal talk, even if in the final analysis possible worlds are analyzed in terms of modality rather than the other way around.

The problem, then, is the following.[7]

---

[5] Kripke, *Naming and Necessity* (Cambridge: Harvard, 1980); Keith Donnellan, "Proper Names and Identifying Descriptions," in Donald Davidson and Gilbert Harman, eds., *Semantics of Natural Language* (Dordrecht: Reidel, 1972), pp. 356–79; and David Kaplan, "Demonstratives," in Joseph Almog, John Perry, and Howard Wettstein, eds., *Themes from Kaplan* (New York: Oxford, 1989), pp. 481–564.

[6] A presentist might construct surrogates for past times out of materials existing in the present—for example, from propositions; see, for example, Prior, p. 138. But the tenses will be used in constructing these surrogate times; hence the time surrogates will not be available for use in *analyzing* the tenses. The issues here are parallel to those which arise for actualists who construct possible-world surrogates from actually existing abstract entities—see Robert Merrihew Adams, "Theories of Actuality," *Noûs*, VIII (1974): 211-31; Alvin Plantinga, "Actualism and Possible Worlds," *Theoria*, XLII (1976): 139-60; Robert Stalnaker, "Possible Worlds," *Noûs*, X (1976): 65-75; and Lewis, *On the Plurality of Worlds* (New York: Blackwell, 1986), chapter three.

[7] The problem need not be stated with names; consider, for example, 'Some current philosophers admire ancient Greek astronomers'. For a recent discussion of this problem, see Bigelow.

(A) David Lewis admires Frank Ramsey.

(A) seems true, but what is the presentist's translation? Even waiving objections to descriptivism about names, (A) is still problematic. The following attempted paraphrase fails to be true because, according to the presentist, the description 'The inventor of the best-system analysis of lawhood' is (currently) nonreferring:

($A_P$1) The inventor of modal realism admires the inventor of the best-system analysis of lawhood.

Adding a past-tense operator does not help:

($A_P$2) WAS (The inventor of modal realism admires the inventor of the best-system analysis of lawhood).

This is false because Ramsey and Lewis never existed at the same time. ($A_P$2) is true only if the component sentence 'The inventor of modal realism admires the inventor of the best-system analysis of lawhood' is true at some past time, and that component sentence is true at a time only if both definite descriptions refer to objects then. Each of the following attempts is better:

($A_P$3) The person *x* that is the inventor of modal realism is such that WAS (*x* admires the inventor of the best-system analysis of lawhood).

($A_P$4) WAS (The person *x* that is the inventor of the best-system analysis of lawhood is such that NOW (the inventor of modal realism admires *x*)).

('NOW' is another tense operator, analogous to the modal operator 'ACTUALLY'.) In neither case is there a problem with the descriptions referring, since the time of evaluation shifts, in each case, between the occurrence of the description for Lewis and the description for Ramsey. But a problem remains: the atomic formulas ascribing the admiration relation seem false (with respect to the times in question). Roughly, in ($A_P$3) it is asserted that at some past time at which Ramsey existed, Lewis admired him, even though Lewis did not exist then; and in ($A_P$4) the assertion is that it is now true that Lewis and Ramsey stand in the admiration relation, despite Ramsey's current nonexistence.[8]

---

[8] Notice that presentism does not on its own rule out the truth of ($A_P$3) and ($A_P$4); for that we need a stronger claim which might be called *serious presentism*, by analogy with Alvin Plantinga's term *serious actualism*—"Replies to My Colleagues," in James Tomberlin and Peter van Inwagen eds., *Alvin Plantinga* (Dordrecht: Reidel, 1985), pp. 316-23 and 345-49. Since presentists should accept the truth of 'WAS (there is an *x* such that NOW (*x* does not exist))', presentism does not on its own

## III. UNDERLYING TRUTHS

My defense of presentism against the objections of the previous section will be subject to the following constraint: the presentist should not *completely* reject ordinary talk and thought. The presentist should salvage *something* from what we commonly say. Not just because we say it; we should take everyday talk seriously because we typically have decent evidence for what we say. A presentist who completely rejected masses of ordinary talk as just being confused would be a quite radical skeptic.

But I do *not* assume that the presentist needs to demonstrate that ordinary talk is *true*. I follow the lead of Ned Markosian, who in a recent paper suggests biting the bullet and admitting that sentences naming merely past individuals or ascribing cross-time relations are not true (*op. cit.*). Markosian's suggestion is that while such sentences are not true, we can explain their *appeal* by noting that there are related truths with which they are commonly confused.

In light of Kripkean considerations, the qualitative sentence ($L_P$) is not synonymous with (L). ($L_P$) is semantically close enough to (L), however, that it could plausibly be said to be confuseable with (L). (L) *seems* true to us, Markosian would say, because we do not adequately distinguish it from ($L_P$), which *is* true. Consider next the case of cross-time relations. Markosian's example is:

(G) There was a great grandfather of Ned.

A truth with which (G) might be confused is:

($G_M$) WAS {there is an $x$ that is a father of Ned & WAS [there is a $y$ that is a father of $x$ & WAS (there is a $z$ that is a father of $y$)]}.

---

rule out the truth of the parallel ⌜WAS (there exists an $x$ such that $x$ is a dinosaur, and NOW ($x$ is a dinosaur))'; for that, we need serious presentism, which may be formulated as the conjunction of presentism and the additional claim that "positive" atomic formulas, like '$x$ is a dinosaur' and '$x$ admires $y$', can never be true with respect to times at which the referents of the names and variables contained do not exist (more carefully: for every positive atomic formula $\phi$ with variables $x_1 \ldots x_n$ free, the following is true: ⌜ALWAYS, for all $x_1, \ldots$ALWAYS, for all $x_n$, ALWAYS: if $\phi$ then $x_1 \ldots x_n$ all (currently) exist⌝). I shall assume that the presentism to be defended is serious presentism.

I shall also assume the unacceptability of two other views which, if true, could help with some of the difficulties in the text: (i) using temporal analogues of "actually" operators that can be indexed to occurrences of tense operators and occurrences of variables and names within atomic formulas—see Graham Forbes, *The Metaphysics of Modality* (New York: Oxford, 1985), pp 90-93; (ii) accepting the current existence of uninstantiated individual essences of merely past individuals; for the modal analogue of this view, see Plantinga, "Actualism and Possible Worlds," section 5; for criticism, see Robert M. Adams, "Actualism and Thisness," *Synthese*, XLIX (1981): 3-41; and Kit Fine, "Plantinga on the Reduction of Possibilist Discourse," in van Inwagen and Tomberlin, pp. 145-86

The strategy here is to find sequences of temporally overlapping objects underlying ascriptions of cross-time relations. And for cases like (A), Markosian's suggestion would be the following:

> ($A_M$) There are various properties $p_1$-$p_n$ such that (i) Lewis associates $p_1$-$p_n$ with the name 'Ramsey', (ii) WAS (there is an object that has $p_1$-$p_n$ and is called 'Ramsey'), and (iii) if there were an object that had $p_1$-$p_n$, Lewis would admire it.

I believe that the general idea of giving up on paraphrasing problematic sentences is a fruitful one. The core claim here is that, even if we deny that the problematic sentences are true, we can still accord them some positive status, and thus not saddle ourselves with an implausible skepticism. But I do not follow Markosian in identifying this positive status with *confuseability with a truth.* I shall seek *underlying* truths for claims like (G), (A), and (L), but I shall not claim that speakers confuse those underlying truths with the propositions they assert.

The problem with confuseability as the status for the underliers is that the underliers are going to need to get pretty elaborate, to the point of its being implausible that we confuse them with anything. ($G_M$) needs to be complicated, for example, since Ned's great grandfather might have died and been cremated before having children, Ned's grandfather being created using the great-grandfather's frozen sperm cells. This sort of possibility would only be accounted for by a complicated tensed sentence mentioning sperm and egg cells, which no one would confuse with (G). In the case of ($A_M$), the problem is that an implausible descriptivism about admiring is presupposed. The properties Lewis associates with 'Ramsey' may not have been had by Ramsey; or, there might be someone else who also had the same properties that Lewis attributes to Ramsey, and who was named 'Ramsey', who Lewis does not admire. What attaches Lewis's admiration to Ramsey himself is at least in part, I think, some kind of causal connection between Lewis and Ramsey, and not a matter of properties that Lewis attributes to Ramsey. But filling in the details properly here will, I suspect, require a quite complicated underlier, unlikely to be confused with (A). My suspicion is that these sorts of complications will quite generally need to be made in Markosian's underlying truths.

In place of seeking underliers that are confuseable with the originals, I suggest a more modest goal of seeking truths that suffice for the claim in question to be, if not true, then at least *quasi true,* as I shall say. I shall introduce the notion of quasi truth informally at first, and then give a more precise characterization. The working idea of a quasi-true sentence is one that, *philosophical niceties aside*, is true. Put a second way, a sentence is quasi true if the world is *similar enough* to the way it would have to be for the sentence to be genuinely true. A third characterization specifies quasi truth by the role I want it to play in my defense of presentism. To remain plausible, presentism should not require us to alter drastically our beliefs about the past; giving up on our ordinary beliefs being true, but retaining belief in their quasi truth, is intended to be sufficiently nondrastic alteration.

Let us look at an example in some detail. Ordinary folks say 'Abraham Lincoln was tall'. The presentist's reconstruction of the past renders this sentence almost, but not quite true. There are no true singular propositions *about* Lincoln since such propositions do not exist; what is true instead is a network of tensed propositions. In this network are included various descriptive facts, such as the fact that there was someone named 'Abraham Lincoln', who was president of the United States, signed the Emancipation Proclamation, and so on, and who was tall. Antidescriptivist reasons for thinking that such facts are not sufficient for the truth of the sentence in question are familiar. But the presentist can get much closer. For consider other facts that the antidescriptivist thinks are relevant to the truth of the sentence, such as facts about the causal chain connecting Lincoln to current uses of 'Lincoln'. Such facts have their presentist analogues: the network will include a variety of tensed facts specifying an initial baptism of 'Lincoln', subsequent utterances of that name, and so on.

In the network, in fact, can be included everything one could say in the presentist's tensed language about the relevant bit of the world *at a subatomic level*. Thus, the presentist can provide a sort of supervenience base for the sentence 'Abraham Lincoln was tall'. Not in the usual sense, for according to usual understandings of supervenience, the existence of a supervenience base for a sentence renders that sentence true.[9] But the presentist can provide what would be a supervenience base if presentism were false–or better, if eternalism were true. If there is such a "quasi-supervenience base" for a sentence *S*–to a first approximation, a true proposition *P* that would have been true and entailed the truth of *S*, if eternalism were true[10]–then I shall call that proposition *P* an ***underlying truth*** for *S*, and shall call S *quasi true*. And finally, I say that the presentist sufficiently discharges her obligation to "common sense" if she can show that ordinary utterances about the past are quasi true.

---

[9] On the most common supervenience terminology, supervenience applies to sets of properties (or predicates) not sentences (or propositions). See, for example, Jaegwon Kim, "Supervenience and Supervenient Causation," *Southern Journal of Philosophy*, XXII (The Spindel Conference Supplement, 1984): 45-56; and "Concepts of Supervenience," *Philosophy and Phenomenological Research*, XLV (1984): 153-76.

This characterization of quasi truth, I believe, fits my initial gloss of quasi truth as "truth, philosophical niceties aside," and "similar enough to the truth." Moreover, I also think the notion of quasi truth can play the role I have prepared for it. I said earlier that a presentist who completely rejected ordinary talk about the past would be committed to an implausible skepticism, because we typically have good evidence for our claims about the past. But if the presentist can maintain that ordinary talk about the past is quasi true, then this skepticism is avoided. Surely, ordinary empirical evidence does not favor the truth of 'Lincoln was tall' *over its quasi truth*. The reason is that ordinary empirical evidence seems qualitative.[11] Ordinary empirical inquiry will justify a belief that there was an object with certain qualitative features. One needs philosophical argument, which goes beyond ordinary empirical justification, to support the further claim that a singular proposition about a particular past object exists and is true. Thus, my presentist attributes to sentences about the past a status that is supported by ordinary evidence. Moreover, my presentist says that the past is quite similar to what would be required for the truth of our utterances about it. Objectionable skepticism is thereby avoided.

Once we give up confuseability as the mark of an underlier, we need an alternate psychological explanation of why sentences like (G) appear true. But an explanation is readily found: in the ontologically unscrupulous nature of natural language. In ordinary life we do not speak as if presentism is true: we quantify freely over merely past objects, and we are taught to talk as if such objects exist. Similarly, in ordinary life we quantify freely over nonactual objects, and over abstracta, without thinking very hard about whether such objects exist; we say that there are many ways to skin a cat, without worrying about what, exactly, a way is. This is psychological explanation enough.

---

[10] Worries about the definition of quasi truth, including the worry that conditionals of this sort are invariably vacuously true, are addressed in section IV below

[11] In saying that ordinary empirical evidence is qualitative, I do not mean to deny that ordinary empirical inquiry typically results in justified belief in singular propositions. I grant that in typical cases, given good evidence, a justified belief results in a proposition about a particular individual, if that proposition exists It does not follow that the evidence justifies the conclusion that the proposition exists; it could simply be the case that, *if* the proposition exists, it is justified.

It is no accident that ordinary language quantifies so readily. It is convenient to quantify over things you do not really believe in, if there is a way to pass back and forth between quasi truths and real truths. Even if individual speakers cannot perform these conversions, it is easy to see how a group of people could evolve a practice of free-wheeling quantification: the talk would be useful, whether or not anyone is capable of actually doing the conversions. Suppose that, in fact, there are no such things as propositions, by which I mean that there are no entities that are capable of playing the role propositions are supposed to play. (This might be so, if a very strong form of nominalism held.) Sentences involving apparent quantification over propositions–for example, 'Starbuck believes everything Boomer says'–would nevertheless be quasi true (provided we refine our definition of quasi truth to eliminate the assumption of propositions–see section IV below). The underlying truths might be facts about Starbuck's brain, and his causal relations to Boomer's utterances. Despite the fact that no one thinks about these underlying truths, and no one can "translate" sentences about Starbuck's beliefs into sentences about these underliers, it will be useful to talk about Starbuck's beliefs in this way in virtue of the *truth* of the underliers. Since the underlying truth would be a supervenience base for the sentence 'Starbuck believes everything Boomer says' if propositions existed, the quasi truth of this sentence will have the same implications for other matters, such as Starbuck's behavior, as would the truth of the sentence. The sentence will be just as useful for explaining behavior and other matters as it would be if it were genuinely true (more on this example below).

The strategy, then, of seeking nonsynonymous underliers for truths about the past requires only that the underliers suffice for quasi truth, and not truth; and moreover, there is no requirement that ordinary people have thought of the underliers. I like this strategy for dealing with ontological commitment, and not just in the case of presentism. I discuss other applications of the strategy below; for the remainder of this section, I would like to explore just how far the method will take us in the defense of presentism.

First, once we give up on synonymy, we need no *general* paraphrase technique that will give "non-ad-hoc paraphrases" for truths about the past. For the objection that a proposed underlier is too complicated, or is ad hoc, seems to be based on the assumption that the underlier is intended to be synonymous with the original.

Second, given the way I have "lowered the bar" and required of ordinary utterances mere quasi supervenience on presentist facts, the presentist can answer many challenges all at once. For example: in the case of cross-time relations, all *internal* relations are immediately rendered unproblematic. Internal relations are those which supervene on the intrinsic properties of the relata. But intrinsic properties of objects at times may be captured in the tensed language of the presentist; cross-time internal relations thus quasi supervene on facts acceptable to the presentist. Consider, for example, the assertion that there is someone who is exactly as tall as some ancient Greek philosopher. *As-tall-as* is an internal relation, the holding of which supervenes on the heights of the relata. The *nonpresentist* will agree that the truth of any cross-time ascription of this relation will be entailed by a true tensed proposition asserting the heights of the involved objects; in the present case such a proposition might be the proposition that '*There is a person who is five feet tall exactly, and it was the case over 2000 years ago that there exists a Greek philosopher who is exactly five feet tall*'. Given presentism, this proposition does not entail the truth of 'There is someone who is exactly as tall as some ancient Greek philosopher'; but it counts as an underlying truth for this sentence since the entailment holds if eternalism is true. Therefore, the sentence is quasi true.

External relations—for example, spatiotemporal relations—do not supervene on the intrinsic natures of their relata. A presentist, therefore, will need to find a quasi-supervenience basis for all cross-time external relations (and also cross-time relations that are neither internal nor external[12]). But it is a reasonable hypothesis that, as the nonpresentist would put it, all relations supervene (globally) on the totality of facts about (i) where and when intrinsic properties are instantiated, and (ii) nomological matters, including causal relations and laws of nature. Indeed, it is a reasonable hypothesis that all facts whatsoever supervene on this basis.[13] So if a presentist can find a quasi-supervenience basis for nomological and spatiotemporal facts, then she will have found a quasi-supervenience basis for all facts, and hence will have solved, in one fell swoop, all ontological problems for presentism of the sort considered in this paper: any utterance deemed true by the nonpresentist will turn out quasi true. I consider classes (i) and (ii) of facts in turn.

---

[12] For this taxonomy of relations, see Lewis, *On the Plurality of Worlds*, p. 62.

[13] Haecceitists will disagree here. At this point, I think the presentist needs to take a stand and reject haecceitism, at least about nonpresent objects; for example, a presentist must deny that, if eternalism were true, 'Lincoln was tall' could differ in truth value between possible worlds that are alike with respect to qualitative tensed facts and with respect to singular propositions about present objects.

Using tensed sentences of the form 'It WAS/WILL BE the case *n* units of time ago/hence that there is an object with intrinsic property *F*', the presentist can provide a quasi-supervenience basis for many claims regarding the instantiation of intrinsic properties at past and future times. What is less clear is that the full range of claims of *spatio* temporal property instantiation in science and everyday life can be given a quasi-supervenience basis. As above mentioned (footnote 1), I am ignoring the well-known apparent conflict between presentism and special relativity. But even setting this aside, there is a question about how a presentist will ground claims that specify both that a property was (or will be) instantiated, and also, roughly, *where* that property was instantiated. The problem would be manageable if we were willing to accept a Newtonian conception of substantival space persisting through time, together with its associated notion of absolute rest and position, for then tensed claims of the following form could serve as underlying truths: 'It WAS/WILL BE the case *n* units of time ago/hence that there is an object at place *p* with intrinsic property *F*'. But without this assumption, it is at least prima facie difficult to see how a presentist could provide underlying truths for certain statements involving the comparison of positions at different times—for example, the claim that a certain particle has been in a state of inertial motion throughout a certain period of time.[14] I shall say no more about this matter here, save that it is a challenge that, to my knowledge, presentists have not yet adequately faced.

If the spatiotemporal pattern of instantiation of intrinsic properties can be given a quasi-supervenience basis, the defense of presentism then reduces to the problem of showing that causation and laws of nature quasi supervene on the totality of tensed facts that the presentist accepts. First note that, if Humean supervenience[15] is true, this problem can indeed be solved. According to Humean supervenience (as a nonpresentist would put it, anyway), the totality of facts about the instantiation of local qualities throughout space-time settles all other facts, including nomological facts.[16] If the difficulty of the previous paragraph can be answered, every case that a nonpresentist would describe as the instantiation of a local quality at a space-time point has a presentist analogue: a tensed fact about the instantiation of a local quality. Hence, if Humean supervenience is true, these tensed facts will form a quasi-supervenience basis for facts about everything, and so for facts about laws.

---

[14] Note that we do not need to make sense of absolute rest in order to make sense of inertial motion—in both Minkowski and neo-Newtonian space-times, the latter but not the former is well defined.

[15] See Lewis, *Philosophical Papers, Volume II* (New York: Oxford, 1986), pp. ix-xvii, and "Humean Supervenience Debugged," *Mind*, ciii (1994): 473-90.

[16] Theories of lawhood that are consistent with Humean supervenience include Lewis's version of Ramsey's best-system theory of laws (the most recent exposition is in "Humean Supervenience Debugged"), and the traditional regularity theory. For a critical discussion of the regularity theory, see Armstrong, *What Is a Law of Nature?* (New York: Cambridge, 1983), part 1. Theories of causation that explain causation in terms of laws will be consistent with Humean supervenience if one of these Humean accounts of lawhood is true.

Suppose, on the other hand, that Humean supervenience is false. The leading non-Humean view of laws of nature–defended by Fred Dretske, D. M. Armstrong, and Michael Tooley[17]–is the view that laws of nature are relations among universals. On this view, we have a law that *Fs* are *Gs* if and only if the "necessitation" relation holds between the universals *F* and *G*. Laws would be unproblematic for presentists on this view if the relevant universals all currently exist, for then statements of lawhood could be straightforwardly true–quasi truth would not be needed. But on Armstrong's version of the theory, universals do not exist unless they are instantiated; for the presentist, this formula becomes: a universal exists only if it is *currently* instantiated.[18] This means trouble if some universals involved in the laws used to be instantiated but are no longer. Imagine that a certain sort of subatomic particle comes into existence only under very extreme conditions, and imagine that these conditions were only created once, in the past.[19] A presentist might try appealing to descriptive propositions about these merely past universals, but this works only if universals have essences that can be captured by such descriptive means.

Turning next to causation: on many views, facts about causation supervene on facts about laws plus the instantiation of qualitative properties and single-time relations. If this is true, then attributions of causation will be quasi true for the presentist provided that attributions of laws are (since instantiation of qualitative properties and single-time relations can be captured by the presentist's tensed claims). But problems will arise if *singularism* about causation is true; if, that is, causal relations hold independently of facts about laws. The problem is particularly acute if the causal relation can hold between temporally distant events.[20] What would be the underlying truths for this sort of causation at a temporal distance?

---

[17] Dretske, "Laws of Nature," *Philosophy of Science*, XLV (1977). 248-68, Armstrong (ibid.), Tooley, *Causation: A Realist Approach* (New York: Oxford, 1987).

[18] For Tooley, universals are transcendent and hence can exist uninstantiated; the problem discussed in the text would therefore not arise. See Tooley, pp. 72-75 and section 3.2.

[19] Compare Tooley, pp. 72-73, who discusses the parallel case of a universal that never is actually instantiated, but might have been if certain conditions had arisen.

[20] If causally related entities were always connected by a chain of temporally overlapping causally related events, then the presentist might be able to utilize tensed claims describing these chains to provide a quasi-supervenience basis for attributions of causation.

One might provide the underlying truths by accepting a *sentence-operator* account of causation. The fundamental locution on this view is not 'event $e_1$ causes event $e_2$', but rather involves a two-place tense operator 'BECAUSE $\phi$, it WILL be the case $n$ units of time hence that $\psi$', where $\phi$ and $\psi$ are filled in with sentences.[21] Thus, if someone's current happiness is caused by someone's eating dinner an hour ago, the underlier is the following:

> ONE HOUR AGO (BECAUSE someone is eating dinner, it WILL be the case in an hour that someone is happy).

This seems to be the best the presentist can do here, though the approach places some limits on possibilities for causation. Imagine a world where objects pop out of existence, causing distinct objects to pop into existence an hour later, and suppose that balls *A* and *B* disappear, and an hour later, balls *C* and *D* appear. Which of the two balls appearing were caused by which of the first two balls? It seems that there are two distinct possibilities; *A* could cause *C* while *B* causes *D*, or, on the other hand, *A* could cause *D* while *B* causes *C*. To account for these two possibilities, the presentist must come up with underlying facts that distinguish them; but there appears to be only a single underlying tensed fact:[22]

> ONE HOUR AGO (BECAUSE a ball disappears, one hour hence, a ball WILL appear; and BECAUSE a ball disappears, one hour hence, a ball WILL appear).

The presentist can distinguish these possibilities if there are qualitative differences between the balls in virtue of which the causal relations hold; if balls *A* and *C* are red and *B* and *D* are blue, then in one world we can say that a red ball's disappearing causes a red ball's appearing, whereas in the other world a red ball's disappearing causes a blue ball's appearing. But one would have thought that this sort of scenario could have occurred with duplicate balls, or, more exotically, with balls that, by virtue of symmetry in their worlds, have exactly the same qualitative features, both relational and intrinsic. Moreover, one would have thought that causal differences of this sort could obtain *without* obtaining in virtue of the qualitative differences of the involved objects. These are possibilities that can only be admitted by a nonpresentist.

---

[21] For similar proposals about presentist causation (without the present reservations), compare Adams, "Time and Thisness," in P. French, T. Uehling, and H. Wettstein, eds., *Midwest Studies in Philosophy, Volume XI* (Minneapolis: Minnesota UP, 1986), pp. 315-29, here p. 321; Bigelow, pp. 39-43; and Dean W. Zimmerman, "Chisholm and the Essences of Events," in Lewis Edwin Hahn, ed., *The Philosophy of Roderick M. Chisholm* (Chicago: Open Court, 1997), pp. 90-92.

[22] We might try to make the caused state of affairs in the underlying fact *de re*:

> There currently exist two balls, *C* and *D*, and ONE HOUR AGO (there exists an *x* and a *y*, and BECAUSE *x* disappears, one hour hence, *C* WILL appear; and BECAUSE *y* disappears, one hour hence, *D* WILL appear).

But the truth of this would require the falsity of serious presentism, a move we are currently trying to avoid; see footnote 8.

This discussion of spatiotemporal property instantiation, laws, and causation shows that the present method for defending presentism is not a blank check. My claim has been that presentism remains plausible if, or to the extent that, ordinary statements about the past can be shown to be quasi true. Establishing quasi truth in effect amounts to showing that intuitively distinct possibilities can be distinguished on the presentist's terms, that is, using the presentist's tensed language and assuming the nonexistence of nonpresent things. And it is an open question to what extent this can be done. Indeed, given certain assumptions about laws, causation, and related matters, the presentist cannot show that all ordinary beliefs about causation are quasi true. The price to pay for presentism, therefore, is rejecting these assumptions; the price for the assumptions is rejecting presentism.

## IV MORE ON THE NOTION OF QUASI TRUTH

I said in the previous section that a sentence is quasi true if there is some true proposition that, if eternalism were true, would be true, and would entail the truth of that sentence.[23] There are several questions to be asked about how this definition is to be understood. First, since eternalism seems to be the sort of proposition that is metaphysically necessarily false if actually false, the conditional beginning with 'if eternalism were true' cannot be understood in such a way that it is vacuously true just in virtue of having a metaphysically impossible antecedent. I do not regard this as a great obstacle to understanding the conditional. Even if the denial of presentism is metaphysically impossible, in some broader sense it is not impossible; it is not impossible in the way that 'It is raining and also it is not raining', and 'Some bachelors are married' are impossible. For lack of a better term, I shall call this broader sense of possibility *logical possibility*. I think it is plain that we do make nonvacuous sense of counterfactual conditionals with metaphysically impossible but logically possible antecedents, and I propose to make sense of one myself in giving my definition.[24]

---

[23] It is important that the definition require that, if eternalism were true, the underlier would be *true*, as well as that it would entail *S*; otherwise, every sentence, *S*, would turn out quasi true: the underlier would be the proposition that *either eternalism is false or S is true*. Thanks to John Hawthorne for a helpful discussion of this matter.

Second, we must ask about the strength of 'entails'. The underliers I have in mind do not *logically* entail the target sentences, since, the presentist supposes, presentism is actually true but the target sentence is not. The right strength seems to be that of metaphysical entailment, by which I mean that in no world that would have been metaphysically possible (if eternalism were true) is the underlier true and the sentence false. Thus, the definition reads: *S* is quasi true if there is some true proposition that would have been true and *metaphysically* entailed *S*'s truth, if eternalism had been true.

There is finally the question of the ontology and ideology required by my definition of quasi truth. The commitment to modal talk via supervenience is at the heart of my proposal, and presumably ineliminable, though it should be noted that this need not require a commitment to unactualized possibilia; one could take modal notions as primitive, or reduce modality in some way that does not require possibilia. A more worrisome feature of the definition is its assumption that underlying truths are propositions. This assumption might be unwelcome, most notably if the method of quasi truth is to be used to eliminate commitment to propositions themselves. One remedy would be to utilize sentences rather than propositions. This could succeed even if sentences are unsuited to *generally* replace propositions, for many of the familiar limitations of sentences do not affect my use of underliers (for example, the underliers do not need to be objects of belief or semantic values of "that clauses"). But a worry persists: we do not want underlying truths to be limited to those expressible in human languages. There are various fixes, none perfectly satisfactory; here are two. (1) Let underliers be sentences in a "Lagadonian" language, in which sparse universals, whether or not humans know of them, are used as predicates denoting (expressing) themselves.[illegible] Cost: commitment to sparse universals. (2) Since we have no need for false underliers, let the underlying truths be facts, rather than propositions. Cost: commitment to facts.

---

[24] On this topic, see Jeffrey Goodman, "Extended Ersatz Realism" (unpublished); William Lycan, *Modality and Meaning* (Boston: Kluwer, 1994), pp. 38-39; and Takashi Yagasawa, "Beyond Possible Worlds," *Philosophical Studies*, LIII (1987): 175-204. The counterfactual conditional *may* not be required; the conditional could perhaps be one of entailment where the modality is more strict than metaphysical necessity, but not so strict as purely syntactic or model-theoretic entailment.

A skeptic might allow *some* counterpossible conditionals but balk at the rich array of counterpossibles I need–conditionals expressing complex supervenience relations that would hold if certain metaphysically impossible theses were true. While I cannot give a theory of the truth conditions of the counterpossibles I need, I think it can be seen that the worry is unfounded. Whatever one thinks these truth conditions are, exactly–perhaps conditions referring to conventions, or causal or logical facts–the truth makers required for the modal statements I need are presumably available whether or not presentism is true. Thus, the present case is quite different from well-known problematic uses of counterfactuals where the truth makers appear to be missing–counterfactuals that "float" on nothing (see, for example, Armstrong, p. 31).

## V TRUTH AFTER ALL?

I have said that a presentist can reconcile a restrictive ontology with freely quantifying natural language and belief by retreating to quasi truth. But might quasi truth actually suffice for truth?

"Objects are in contact only if there is absolutely no space between them." That is what we would have said before the conception of matter given to us by classical physics; indeed, we would have regarded this as being definitional of contact. We know now that nothing satisfies the definition; must we conclude that earlier folk never truly ascribed the predicate 'contact'? A common response is that since the world was *near enough* to the folk's definition of 'contact', their ascriptions of contact were true. Despite the fact that the folk would have vehemently adhered to their original "definition," their word 'contact' expressed a relation which held in paradigm cases of contact, a relation which perhaps involves lack of visible separation, resistance to further smashing together, and so on. Lewis describes this view about content as follows:

> It's an old story. Maybe nothing could perfectly deserve the name "sensation" unless it were infallibly introspective; or the name "simultaneity" unless it were a frame-independent equivalence relation; or the name "value" unless it couldn't possibly fail to attract anyone who was well acquainted with it. If so, then there are no perfect deservers of these names to be had. But it would be silly to lose our Moorings and deny that there existed any such things as sensations, simultaneity, and values. In each case, an imperfect candidate may deserve the name quite well enough.[illegible]

If a sentence is quasi true, then the world is fairly similar to the way it would need to be for the sentence to be true, since there is a true proposition that would be true and would suffice for the sentence's truth, if eternalism were true. On the view of content in question, would this make the sentence in question true after all?

The question may not have a definite answer. Surely, there is no sharp line dividing "imperfect candidates" from near misses. Never-

[illegible] See Lewis, *On the Plurality of Worlds*, section 3.2

[illegible] "Humean Supervenience Debugged," p. 489

theless, I think there is reason to doubt that quasi truths are truths. We should distinguish between candidates for being expressed by sentences and candidates for being expressed by subsentential expressions. If there is an imperfect but good enough candidate for being referred to by the predicate 'contact', it is plausible to say that the sentence 'Some things come into contact with others' is true. This is less plausible when there is no candidate for the predicate, and only a candidate for the whole sentence; the reason is that there is some pressure, admittedly defeasible, to respect the structure of a sentence in assigning it content. The sentence is, syntactically, a quantified sentence saying that there are objects of a certain type. It seems right to say that this sentence is true only if there really are two objects that stand in the relation of contact.

Call an interpretation of a language *weakly devious* if it respects the logical structure of the language (in other words, gives a conjunctive semantics for syntactic conjunctions, a quantificational semantics for syntactic quantificational sentences, and so on), but "reinterprets" (along the lines of 'contact') some subsentential expressions such as predicates, names, functors, and the like. And call an interpretation *strongly devious* if, and to the extent that, it does not respect the syntactic structure of the language in this way. I am suggesting that, other things being equal, strongly devious interpretations provide worse candidates for reference and meaning than do weakly devious interpretations. The truth of this principle (defeasibly) counts against the truth of the presentist's quasi truths, since the underlying truths I have imagined do not structurally match the sentences in question. The sentence 'There was a Greek philosopher who is exactly the same height as someone currently existing' is, syntactically, the result of applying existential quantifiers to an atomic formula; but the proposed underlying truth is expressed by the conjunction of two tensed existentially quantified sentences: 'There is someone who is exactly five feet tall', and 'WAS (there is someone who is a Greek philosopher, and who is exactly five feet tall)'.

There is a further reason why this underlying truth should not count as a devious meaning for the sentence 'There was a Greek philosopher who is exactly the same height as someone currently existing': it would only be plausible as a *sufficient* condition for the truth of the sentence in question. One could, of course, utilize instead a disjunction of all the underliers, which would itself count as an underlying truth for the sentence, and would be a more plausible candidate for being the devious meaning of the sentence. But there is a further obstacle: I think there is reason to prefer *simple* strongly devious interpretations over extremely complex ones. This principle, if granted, would provide more reason to claim that the presentist's quasi truths are *merely* quasi true.[27]

## VI. OTHER DESERT LANDSCAPES

Presentists are not the only ones who face the dilemma with which we began. Nonpresent objects are not the only dubious things over which natural language and thought quantify. I shall now explore the quasi-truth defense of some restrictive ontologies other than presentism.[28] But first let me point out that the defense will not work for eliminating commitment to theoretical entities in science—electrons, for example. No one who believes in electrons thinks that facts about electrons supervene on other facts—observable facts, for example. Our evidence for claims about electrons admittedly comes from observation, but the familiar fact that theory is underdetermined by data shows that it would be *possible* for the same observational data to be caused by many possible microconfigurations. Observable facts, therefore, do not form a quasi-supervenience base for claims about electrons, and so if there are in fact no electrons, claims about them will not be quasi true. This is, I take it, a desirable feature of the quasi-truth defense, namely, that it brings out an asymmetry between *scientific* and *philosophical* ontological commitment.[29]

---

[27] In a forthcoming paper, "The Ersatz Pluriverse," I provide a strongly devious reinterpretation of sentences apparently quantifying over possibilia, but it is a simple strongly devious reinterpretation, and it is not *too* strongly devious, and so, I think, is still plausible as preserving the truth of such talk. But if I am wrong about this, I would not mind accepting that talk about possibilia is merely quasi true. Note that we would then need an account of quasi truth for noncontingent subject matter; see footnote 28.

[28] I do not pretend that the present approach provides a completely general haven from commitment to unwanted entities. One case I shall not discuss here is that of mathematical entities. Under the present modal definition of quasi truth, my approach is unsuitable for a defense of nominalism about pure mathematics, and in general for cases involving noncontingent truth. Any proposition trivially counts as an underlying truth for sentences of pure mathematics simply because of the fact that, if mathematical entities existed, then any true proposition would be metaphysically necessarily sufficient for all mathematical propositions. Defending desert landscapes is not *that* easy! The presence of such "underlying truths" would not make the world similar to how it would have to have been for the mathematical sentences to be true; and their presence would not make platonistic assertions 'True, philosophical niceties aside.' As for applied mathematics, the quasi-truth defense may well be appropriate, but it seems glib to appeal to quasi truth and leave it at that, for it is not immediately obvious that there are indeed nominalistic underlying truths for the statements of applied mathematics. Indeed, one could regard Hartry Field's *Science without Numbers* (Princeton: University Press, 1980) as putatively establishing that applied mathematical claims are quasi true: the representation and uniqueness theorems that Field discusses in chapters 4 and 6-8 would hold if Platonism were true, and thus in that case, the facts captured by the nominalistic axiomatizations Field discusses would be a quasi-supervenience base for platonistic scientific claims.

A case other than presentism where the defense might be employed was noted above: someone who disbelieved in propositions could hold that statements that appear to quantify over propositions–for example, 'Starbuck believes everything Boomer says'–are quasi true; the underlying truths would be facts about Starbuck and Boomer on which the truth of the sentence would supervene if propositions existed. As I have been arguing, it is not implausible to deny that this sort of sentence is true, provided one can show that it is at least quasi true.

One sort of argument for the existence of propositions has therefore been answered, an argument premised on the apparent truth of sentences that appear to quantify over propositions. Another sort of argument, based on considerations of logical form, remains, however.[30] Argument A1 seems plainly valid:

Argument A1:
Starbuck believes everything Boomer says.
Boomer says that *The Galactica* is under attack.
Therefore, Starbuck believes that *The Galactica* is under attack.

But this could only be true if (i) the first premise is interpreted as quantifying over propositions, (ii) the second premise attributes a relation between Boomer and a proposition, and (iii) the conclusion attributes a (different) relation between Starbuck and that same proposition. But if these facts about the logical form of the contained sentences are correct, then, since sentences like the ones in the argument are often true, in particular ones like the second premise which express relations between persons and propositions, there must exist propositions.

---

[29] These considerations also undermine what otherwise might have seemed a promising solution to the problem at the end of section IV of eliminating the commitment to propositions in the definition of quasi truth. The seemingly promising solution is an alternate definition of quasi truth: *S* is quasi true if and only if, had eternalism been true, *S* would have been true. (This would bring my proposal closer to a certain kind of fictionalism–for example, the fictionalism about possibilia defended in Gideon Rosen's "Modal Fictionalism," *Mind*, XCIX (1990): 327-54.) The problem is that this definition cheapens quasi truth by not requiring the *actual* existence of underlying truths. Granted, the conditional 'Had eternalism been true, *S* would have been true', when true, will hold in virtue of the nature of the actual world; but in some cases a sentence might turn out quasi true despite insufficient grounding in actual fact. It would be bogus for someone who disbelieved in subatomic particles, for example, to say that 'There exist carbon atoms' is quasi true, and yet, the conditional 'If there had existed subatomic particles, then there would have existed carbon atoms' is presumably true.

[30] Compare George Bealer, "Universals," this JOURNAL, XC, 1 (January 1993): 5-32, and Peter van Inwagen's "Meta-ontology," *Erkenntnis*, XLVIII (1998): 233-50.

It is always open to a nominalist about propositions to try her hand at paraphrase; and perhaps she can paraphrase the sentences in the argument in a way that preserves its quantificational structure, but in which the entities quantified over are deemed less objectionable than propositions. But I shall proceed on the assumption that this way out will not succeed. In particular, I shall assume that no nominalistically acceptable paraphrase of the first premise can be obtained.

Nominalists might divide about what to say about the second premise and conclusion, however. Unlike the first premise, they contain no overt quantification over propositions. A nominalist might regard these simple propositional attitude sentences as being true, and as not attributing relations between persons and propositions. In this case, the argument would be regarded as invalid, and so unsound. On the other hand, the nominalist might insist that the second premise and conclusion concern propositions, and are therefore false (or lacking in truth value). In this case, the argument would be regarded as valid but unsound. In each case, then, the argument is unsound; but this generates a serious problem for the nominalist. We clearly enjoy inferential success, make useful predictions, and so on, using quantificational language; if we did not, we would have given up using such language long ago. How can this success be explained if arguments like A1 are invariably unsound?

Notice first that the existence of speakers' intuitions that A1 is valid does not provide a strong reason for thinking that it is indeed valid. The nominalist would be perfectly reasonable in claiming that the argument only seems valid because it shares surface form with arguments that *are* valid–for example:

Joe loves every friend of Frank.
Chet is a friend of Frank.
Therefore, Joe loves Chet.

If analytic philosophy has succeeded in anything, it has been in showing the dangers of being misled by the appearances of language (think of 'the round square', 'the average family', and so on).

What the nominalist must do is explain how we could enjoy inferential success in employing arguments like A1, if such arguments are invariably unsound. The answer lies in the fact that, if a sentence is quasi true, this typically will have many of the same entailments as would its truth. The first premise of the argument, that Starbuck believes everything that Boomer says, is quasi true. It therefore has a true underlier, which concerns facts about Starbuck vis-à-vis Boomer, and which would be true and suffice for the truth of the sentence, if propositions existed. There is likewise an underlier for the claim that Boomer says that *The Galactica* is under attack (nominalists divide, recall, on whether or not this underlier actually suffices for the second premise's truth). Now, surely, these underlying truths stand in logical relations that match the logical relations that hold between the premises and conclusion in the argument, or would have held if propositions had existed.[31] That is, the truths underlying the premises surely imply a truth underlying the conclusion—if the world at the microscopic level is sufficient, but for the nonexistence of propositions, for Starbuck believing everything that Boomer says, and for Boomer saying that *The Galactica* is under attack, then the world will be sufficient at a microscopic level (but for the nonexistence of propositions) for Starbuck believing that *The Galactica* is under attack. Thus, the argument has the feature of *quasi validity*: if the premises are quasi true or true, then so will be the conclusion. Since for ordinary purposes quasi truth is as good as truth, it follows that quasi validity is as good for our inferential purposes as validity, and, moreover, that *quasi soundness* (quasi validity plus true or quasi-true premises) is just as good for our inferential purposes as is soundness. Thus, even if arguments of the kind considered here are invariably unsound, they are in many cases quasi sound, and this sufficiently explains the success of our inferential practices. (Notice that I am not saying that ordinary speakers know any of this. The explanation of our inferential success did not depend on our appreciation of any of the above.)

A similar account of other sorts of quantification over dubious entities can be given. We utter sentences which appear true and which quantify over *ways*:

There are many ways in which I could win this chess match.

and we utilize arguments which appear sound, and whose soundness appears to depend on quantification over ways:

Joe has tried every possible way to outsmart Frank.
Wearing a false mustache is a possible way for Joe to outsmart Frank.
Therefore, Joe has tried wearing a false mustache.

---

[31] How might logical relations vary from world to world? Suppose the nominalist says that in the actual world, simple propositional attitude sentences are true, and so do not concern propositions. A1 is therefore invalid. But if there had existed propositions, surely simple propositional attitude sentences would have then concerned them; A1 would then have been valid.

Sentences that quantify over ways are quasi true despite being false, for if there were such things as ways, facts about ways of winning chess matches would supervene on facts about chess and chess players; such facts are therefore, in this wayless world, underlying truths for sentences quantifying over ways. By according sentences quantifying over ways the status of quasi truth, we avoid being absurd skeptics. And in virtue of these underlying truths, arguments of the kind displayed can be useful despite being unsound, for they are often quasi sound.

## VII. CONCLUSION

W.V. Quine[illegible] said that the ontological commitments of a theory are the values of the bound variables in a first-order rendition of that theory. Ordinary language and thought quantifies over all manner of implausibilia. When first-order paraphrases of everyday locutions were available, Quine took that route, but where paraphrase was not forthcoming, Quine was only too ready simply to jettison ordinary thinking, and adopt new, scientific, theories and ways of talking.

Since then, many have concluded that Quine's willingness to "quine" vast realms of ordinary discourse was overzealous, and I agree. But this has led too many contemporary philosophers to postulate abstracta to preserve ordinary talk. We can yearn with Quine for desert landscapes, yet avoid the Quinean mangling of ordinary talk and thought, if we reject his conception of ontological commitment. Ordinary talk and thought, while not our best theory, is a good theory, and should be respected. It is respected well enough if we regard its sentences as being quasi true.

THEODORE SIDER

Syracuse University

---

[illegible] See, for example, "On What There Is," in his *From a Logical Point of View* (New York: Harper and Row, 1953), pp. 1-19.

# REDUCING THERMODYNAMICS TO STATISTICAL MECHANICS: THE CASE OF ENTROPY*

It has become increasingly clear in recent years that the foundations of *statistical mechanics* (SM) are almost as puzzling as–and in some senses are more puzzling than–the foundations of quantum mechanics. Lawrence Sklar's[1] important review of the field presents many different approaches to the foundational problems in SM. Here, the uninitiated encounter a bewildering variety of schemes aimed at explaining why SM works. What strikes one about these projects is that perhaps all of them crucially rely on surprisingly dubious assumptions. Many projects invoke indispensable "rerandomization" posits, some with Hamiltonian dynamics "turning on" and "turning off." Others work only in the limit of infinite particle number and/or time. Still others rely on special dynamical properties that real systems are very unlikely to possess. The arguments supporting these assumptions are typically murky. Perhaps worse, they leave unclear their relevance to the project of providing a mechanical explanation of why classical or quantum-mechanical systems behave thermodynamically. Philosophically minded observers may feel at a loss when they cannot trace the reasoning from a foundational project to an explanation of why (say) particular gases tend to fill their available volumes. I argue here that we are often correct to feel this way. Most of the approaches to the foundations of SM have severed their link with the original foundational project, the project of demonstrating how real mechanical systems can behave thermodynamically.

My argument focuses on two different frameworks in SM: the Gibbsian and the Boltzmannian. I concentrate on their different definitions of entropy, since they underscore the more general conceptual differences between the two approaches. I argue against the prevailing view that the Holy Grail of modern foundational research

---

* Early versions of this paper were presented at Cambridge University's Σ Club, Oxford University's Colloquium in Honour of Robert Weingard, and the University of Hull's Conference on the Mathematical and Conceptual Foundations of Physics. I am grateful to the audiences at all three talks for useful discussion. I also want to thank Guido Bacciagaluppi, Bob Batterman, Joel Lebowitz, Michael Redhead, T.M. Ridderbos, Janneke van Lith and especially Shelly Goldstein and Jos Uffink. Finally, I thank the LSE's Centre for Philosophy of Natural and Social Sciences.

[1] *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics* (New York: Cambridge, 1993).

0022-362X/99/9607/348–73

should be a Gibbsian entropy that strictly mirrors the behavior of the thermodynamic entropy. The search for the Grail has not only been in vain; it has also been misguided. Attempts to find a Gibbs entropy that avoids the so-called "paradox" of the conservation of the fine-grained entropy have proved unsuccessful. Worse, the very goal of these Gibbsian approaches is radically at odds with the potential reduction of thermodynamics to SM. Inasmuch as one seeks a reduction (in some sense) of thermodynamics to SM, one loses her way by pursuing the Grail. The Boltzmann entropy, neglected for much of this century, therefore deserves serious reconsideration in light of the failure of mainstream Gibbsian SM to tackle the above foundational project. The subsequent discussion more or less follows the order of this sequence of claims.

Before beginning, let me happily concede that for the practice of science, Gibbsian SM is usually to be preferred. Since the values of all the entropy functions I discuss agree at equilibrium, my arguments are necessarily philosophical in nature. My topic is the much-ignored *interpretation* of SM. How we interpret SM affects nearly all of the foundational problems plaguing the subject: for example, understanding reversibility, reduction, the role of probability. No less than the notorious problems in quantum mechanics, these problems cry out for philosophical investigation. Indeed, the demand for foundational investigation is especially pressing in SM; for it has long been recognized that, as E. T. Jaynes[2] expresses it, it "is the enormous *conceptual* difficulty of this field which has retarded progress for so long" (*ibid.*, p. 392). Yet philosophers and theoreticians have rarely tackled the subject. As Jos Uffink remarks, this neglect is no better illustrated than by the fact that Paul and Tatiana Ehrenfest's classic 1911 survey article still represents, in many respects, the state of the art.[3]

## I. REDUCTION AND IRREVERSIBILITY

Consider the canonical example of a gas confined by a partition to a section of a container. Remove the partition. In a short span of time, the gas will relax to equilibrium; that is, it will spread evenly throughout the box, possessing a uniform pressure and temperature. What accounts for this regularity? According to thermodynamics, the gas's approach to equilibrium occurs mainly due to the

---

[2] "Gibbs versus Boltzmann Entropies," *American Journal of Physics*, XXXIII (1965): 391-98.

[3] Uffink, "Nought but Molecules in Motion," *Studies in History and Philosophy of Modern Physics*, XXVII (1996): 373 The Ehrenfests's classic article is translated and reprinted as *The Conceptual Foundations of the Statistical Approach in Mechanics* (New York: Dover, 1990)

Second Law. The Second Law states that an extensive state function, the total differential $\delta S = \delta Q / T$, defined only for equilibrium states, is such that $\Delta S \geq \int \delta Q / T$. Loosely put, for realistic systems, this implies that in the spontaneous evolution of a thermally closed system, the entropy can never decrease and that it attains its maximum value for states at equilibrium. We can therefore think of the Second Law as driving the gas to its new, higher entropy equilibrium state. Using this concept of entropy, thermodynamics is able to capture an extraordinary amount of diverse phenomena under one simple law. Remarkably, whether they are gases filling their available volumes, two iron bars in contact coming to the same temperature, or milk mixing in your coffee, they all have an (indirectly) observable property in common: their entropy increases.

Coffee, milk, and iron bars are not elementary individuals in our ontology, of course. Exactly what are is a matter of great controversy. But so long as one believes that quasi-periodic, *time reversal invariant* (TRI) laws govern elementary matter, we can frame our problem in terms of classical mechanics without fear that quantum mechanics will side-step our problem.[4] (The Schrödinger equation is TRI and quasi periodic, so interpretations of quantum mechanics that do not modify or interrupt the Schrödinger evolution will face a problem isomorphic to the one discussed here. Other interpretations may face similar problems, but one must deal with these on a case-by-case basis.) In classical mechanics, the ontology is particulate, and the coordinates and momenta of all of the particles in the system determine the state of a system. Temporal evolution of states is given by Hamilton's equations of motion, $dq_i/dt = \partial H/\partial p_i$ and $dp_i/dt = -\partial H/\partial q_i$. I shall represent the microstates of $N$-particle systems by a point $X \in \Gamma$ and a $2Nf$-dimensional phase space spanned by $Nf$ momenta and $Nf$ configuration axes, where $f$ is the number of degrees of freedom of each particle. As the system evolves through time, this representative point will trace out a trajectory through $\Gamma$.

We thus have two descriptions of our gas: one mechanical and the other thermodynamical. We would like to know how they relate to one another. This kind of problem is quite familiar to philosophers.

---

[4] Roughly put, TRI means (for deterministic systems) that, if we take any nomically possible sequence of states of a system, then the reverse sequence of "temporally reflected" states is dynamically necessary. That the equations are quasi periodic, or recurrent, means the following: for systems confined to a finite volume and a finite energy and for almost all (with respect to the standard measure) microstates $X$ corresponding to these systems, if a system starts in some state $X$, then at some time in its subsequent evolution it will return either to $X$ or to an arbitrarily small neighborhood $\Delta_X$ around $X$.

Philosophers of biology, psychology, and all of the special sciences are busy trying to demonstrate how the properties and concepts used by their science are (not) reducible in some sense to the properties and concepts of "lower level" sciences. The problem is one of intertheoretic reduction.

One might expect the present reduction would be comparatively easy to complete successfully. After all, few take the thermodynamic properties of gases to demand a new ontology. The thermodynamic properties of gases are, broadly speaking, material properties, and we simply need to determine the mechanical properties to which they correspond. The trouble we immediately face, however, is that the properties of fundamental physics are much different from the properties of thermodynamics. The classical mechanical equations of motion are TRI and quasi periodic, yet the Second Law of thermodynamics is not. Focusing on reversibility, Hans Reichenbach[5] puts the problem the following way:

> The elementary processes of statistical thermodynamics, the motions and collisions of molecules, are supposed to be controlled by the laws of classical mechanics and are therefore reversible. The macroprocesses are irreversible, as we know. How can this irreversibility of macroprocesses be reconciled with the reversibility of microprocesses? It is this paradox which the physicist has to solve when he wishes to account for the direction of thermodynamical processes (*ibid.*, p. 109).

Strictly speaking, this is perhaps not the best way to frame the problem. The fallacy of composition threatens. In philosophy of mind, for example, few think that component physical systems of the brain ought to possess the same properties as the brain (for example, intentionality). Similarly, we should not expect aggregates of particles to have the same properties as individual particles. There is no paradox here. In principle, values of momenta and positions exist that will reproduce a thermodynamic system's motion, at least if one is willing to posit conspiratorial initial conditions.

The worry is really about whether thermodynamics can emerge in some natural way from mechanics. Thermodynamic systems–like chairs, tables, and similar systems picked out by our common object language–are nothing more than complicated arrangements of physical properties. Very few would disagree with this. Thermodynamics does not threaten physicalism. In this very weak sense, thermodynamics is *already* "ontologically reduced" to mechanics. Unlike chairs, however, the thermal features of objects play a role in an extraordinarily powerful science: thermodynamics. We have no hope of defining 'chair' in the language of physics—we do not believe it has a simple nondisjunctive mechanical counterpart. The concept is just too coarse. But thermodynamics defines its concepts comparatively precisely, and it is closer to the level of fundamental physics than is our common object language. It is such a powerful tool in organizing a vast range of phenomena, so striking and explanatorily powerful, that we suspect we can find a simple, nondisjunctive counterpart for it at the (statistical) mechanical level. We suspect the thermodynamic entropy is either identical (a posteriori) to this mechanical correlate or at the very least supervenient upon it. Indeed, we *hope* SM has *already* found this correlate. In this stronger, more Nagelian sense of reduction, where the reduced theory is a "definitional extension" of the reducing theory, or at least an approximation to it, the jury is still out concerning the reduction of thermodynamics to SM.

---

[5] *The Direction of Time* (Berkeley: California UP, 1956).

As Reichenbach recognizes, the problem of macroscopic irreversibility is a warning shot for those seeking such a reduction. For as we know from a century's discussion of Boltzmann's notorious H-Theorem, the TRI and quasi periodicy of the dynamics place severe restrictions on any mechanical definition of entropy. Rather than rehash what should be a familiar debate, let me briefly state what I take to be their main point. Consider the following assumptions:

(A) Entropy is a function $S$ of the dynamical variables $X(t)$ of an individual system.
(B) $S(X(t)) = S(X^*(t))$, where '*' indicates a temporal reflection.
(C) The system is closed ($\Gamma$ is bounded).

If (A), (B), and (C) hold, then the TRI of Hamilton's equations implies $S$ cannot increase monotonically for all initial conditions; and if (A) and (C) hold, the quasi periodicy of the solutions to these equations implies $S$ cannot increase monotonically for all time. In short, if $S$ is a function of the dynamical variables of an individual system, then $S$ cannot exhibit monotonic behavior. This captures the essence of Joseph Loschmidt's, Ernest Zermelo's, and Edward Culverwell's objections to the H-Theorem (see Sklar, pp. 39ff.). As we shall see, it is symptomatic of the confusion in the field that the lessons drawn from this implication are so varied.

Perhaps the most common response in the physics literature is to take this implication as effectively a reductio of assumption (A). (Assumption (B) is never challenged; in section III, (C) will be challenged.) The thermodynamic entropy is understood as displaying monotonically increasing behavior; therefore, many reason, the SM analogue of the entropy cannot be a function of the dynamical variables of an individual system. We should instead conceive of entropy in terms of some function of a collection of systems. Clearly, the fact that $S(X(t))$ cannot monotonically increase is compatible with some *other function* defined on an *ensemble* of systems monotonically increasing. The move to ensembles is sometimes advertised as a "way around" the reversibility and recurrence properties of the classical dynamics.[6] Indeed, Jerome Rothstein explicitly treats it as such in a paper aptly entitled "Loschmidt's and Zermelo's Paradoxes Do Not Exist."[7] Noting that Loschmidt and Zermelo's paradoxes apply to individual systems, Rothstein argues that *since* entropy is a function rightly applied to ensembles, the paradoxes do not threaten proofs of entropy increase. With similar reasoning, many circles have come to see denying (A) as the "answer" to the reversibility and recurrence paradoxes. The reader of Sklar's survey will recall, for instance, the denial of (A) as a refrain repeated with the introduction of nearly every approach.

Of course, this foundational problem is not the primary reason for denying (A). The primary reason is that Gibbs's approach to SM denies (A), and the Gibbsian approach dominates the field (which is the reason why Rothstein feels he can use as a premise the claim that entropy is a function on ensembles). Before criticizing the move of avoiding the paradoxes by denying (A), we need to compare the Gibbsian and Boltzmannian approaches to SM.

## II. GIBBS AND BOLTZMANN ENTROPIES

In his famous work of 1902, J. W. Gibbs[8] worries about the paradox of the specific heats, a problem plaguing the field at the time. He famously reacts to it by developing a very abstract approach to the subject, one not so dependent upon the actual constitution of classical systems. When he arrives at his various definitions of entropy, he tentatively offers them as "analogies" to the thermodynamic entropy rather than as mechanical reductions of it.

---

[6] More than denying *A* is necessary to avoid recurrence in quantum mechanics. If a quantum system has a finite number of degrees of freedom, then recurrence among individuals is matched by recurrence *of the ensemble as a whole*. Thus is the lesson of a theorem by S. Ono, *Mem. Fac. Kyushu University*, XI (1949): 125; and (independently) Ian Percival, "Almost Periodicy and the Quantal H Theorem," *Journal Of Mathematical Physics*, II (1961). 235-39. The theorem shows that the expectation value $\langle F\rangle_t$ of any dynamical variable $F$ in an arbitrary Gibbs ensemble is an "almost periodic" function of $t$. This implies that the probability of a particular value of a macroscopic state must come arbitrarily close to its initial value at arbitrarily large times. To avoid this, one usually goes to the so-called "bulk limit," in which $N \rightarrow \infty$ while keeping $N/V$ finite Doing this, however, only serves to aggravate the problems already saddling the denial of (A); see section III.

[7] *Foundations of Physics*, IV (1974): 83-89.

[8] *Elementary Principles in Statistical Mechanics* (New Haven: Yale, 1902; reprint, New York: Dover, 1962).

On Gibbs's approach, one considers a fictitious infinite ensemble consisting of all the microstates compatible with a given macrostate. The macroscopic parameters thus pick out a distribution of possible representative points, which we treat as a fluid in Γ. Upon this ensemble we specify a normalized density function $\rho(p,q,t)$. The mean values of a phase function with respect to $\rho$ describe the system's thermodynamic properties, except for the entropy and temperature. For the entropy, Gibbs chose the expression:

$$(1)\quad S_{FG}(\rho(X)) = -k\int \rho(X)\,[\log \rho(X)]\,d\Gamma$$

where the integral is over Γ and $k$ is Boltzmann's constant. This is the so-called "fine-grained" entropy. For isolated systems, we use the microcanonical probability distribution in (1) and this will match (up to an additive constant) the value for the thermodynamic entropy. (1) will also allow us to define a temperature. Remarkably, using these definitions Gibbs can recover the familiar thermodynamic relations for systems in equilibrium.

Although Gibbs himself originally introduced ensembles as a kind of mathematical trick useful for calculating observed values of macroscopic parameters, influential physicists such as D. Ter Haar and R. C. Tolman[9] dissented from this view. They claimed that the adoption of ensembles is not merely pragmatically desirable but is warranted by the inevitable imprecision of our measurements. Today, it is fair to say that this is the orthodox opinion. Their use in SM is so ingrained that, as D. Lavis[10] expresses it, one needs "a considerable act of will...to avoid all references to ensembles" (*ibid.*, p. 260). Oliver Penrose,[11] for instance, is able to begin a major review of the foundations of SM with the claim that the fundamental problem facing the subject is justifying the choice of the correct ensemble.

How much, if at all, the popularity of Gibbs's methods owes to the denial of (A) as a response to the recurrence and reversibility paradoxes is a historical question I am not competent to judge. Surely, the calculational ease of the Gibbs approach is the primary reason why it dominates other approaches. But Gibbs's worries about the paradox of the specific heats surely struck a chord with many like-minded scientists. Scientists often cited this paradox as a partial justification of instrumentalism and forms of positivism.[12] It would be interesting to know whether these philosophical ideas played any role in the subsequent popularity of Gibbs's theory.

---

[9] Ter Haar, "Foundations of Statistical Mechanics," *Reviews of Modern Physics*, XXVII (1955): 289; and Tolman, *The Principles of Statistical Mechanics* (New York: Oxford, 1938; reprint, New York: Dover, 1979).

[10] "The Role of Statistical Mechanics in Classical Physics," *British Journal for the Philosophy of Science*, XXVIII (1977): 255-79.

[11] "Foundations of Statistical Mechanics," *Reports on the Progress of Physics*, XLII (1979): 1937-2006.

Let us now turn to Ludwig Boltzmann's[13] approach, in particular, his later approach after the H-theorem. In contrast to the Gibbs entropy, we define the Boltzmann entropy $S_B$ for the actual microstate of an individual system. Consider some microstate $X$. $X$ corresponds to a macrostate $M(X)$, which, in turn, is compatible with many different microstates. We wish to determine the relative volume in $\Gamma$ corresponding to all the microstates giving rise to $M$. To accomplish this, we must partition $\Gamma$ into compartments, such that all of the microstates $X$ in a compartment are macroscopically indistinguishable, that is, they share the same thermodynamic features. We do this however we can, but for the ideal gas we use six-dimensional μ-space. Here, we can write the total energy of the system as a sum of one-particle energy functions on μ. The μ-space description is important only for discovering which microstates are macroscopically indistinguishable. Once we determine this, we are then able to associate with each $M$ a certain volume of $\Gamma$, $|\Gamma_M|$, which is the integral of the time invariant Liouville volume element over $\Gamma_M$. Up to additive constants, $S_B$ is then given by:

(2) $S_B(X) = k \log |\Gamma_{M(X)}|$

where $|\Gamma_M|$ is the volume in $\Gamma$ associated with $M$. $S_B$ can be thought of as providing a relative measure of the amount of $\Gamma$ corresponding to the $M$ determined by $X$.

The volume corresponding to a particular macrostate depends, of course, on the macrostate in question. By far the compartment with the greatest volume in Γ is the one each of whose points corresponds to a Maxwellian distribution (in μ-space). When $N$ is large enough for the system to be macroscopic ($\approx 10^{23}$) it occupies almost all of Γ. If we think of a gas in a box, it is easy to see why this is so. There are simply many more ways for a gas to be uniformly distributed in a box, with uniform temperature and pressure, than for a gas to occupy a small corner of the box, with nonuniform temperature and pressure. Almost all of the available Γ, therefore, corresponds to a compartment representing a macrostate in thermal equilibrium. If we adopt the "natural" Euclidean probability metric over Γ, we may say that thermal equilibrium is the most probable state of an individual system. At equilibrium, for macroscopic $N$ at a given energy and volume, $S_B$ is approximately equal to the thermodynamic entropy. In section V, I shall return to our discussion of $S_B$ in order to clarify how it works and how it deals with various difficulties.[14]

---

[12] See Henk de Regt, "Philosophy and the Kinetic Theory of Gases," *The British Journal for the Philosophy of Science*, XLVII (1996): 31-62, and references therein.

[13] See, for instance, *Vorlesungen uber Gastheorie*, Volumes 1 and 2 (Leipzig: Barth, 1896-98)–translated by S. G. Brush as *Lectures on Gas Theory* (New York. Dover, 1964); *Theoretical Physics and Philosophical Problems: Selected Writings*, B. McGuinness, ed. (Dordrecht: Reidel, 1974). For recent work on Boltzmann's theory, see J.L. Lebowitz, "Boltzmann's Entropy and Time's Arrow," *Physics Today* (September 1993)· 32-38, and "Macroscopic Laws, Microscopic Dynamics, Time's Arrow and Boltzmann's Entropy," *Physica* A, CXCIV (1993): 1-27, J. Bricmont, "Science of Chaos or Chaos in Science?" in *The Flight from Science and Reason: Annals of the New York Academy of Science*, DCCLXXV (1996): 131-75; S. Ito, M. Mizutani, and T. Niwa, "On the Time Evolution of the Boltzmann Entropy," *Journal of Mathematics of Kyoto University*, XXVI, 1 (1986): 1-11

What is the relationship between the Boltzmann and Gibbs entropies? As the Ehrenfests and others have pointed out, they take the same values at equilibrium. If the generalized microcanonical ensemble associated with $M$ is defined as $|\Gamma_M|^{-1}$ if $X \in \Gamma_M$ and 0 otherwise, then the two are identical at equilibrium (Lebowitz). This is not surprising, of course, for both approaches essentially seek to calculate the region of motion in Γ available to a system.

Nevertheless, *conceptually* the two entropies are *very* different. Most significantly, $S_B$ is a function of the microstate of an individual system, in contrast to $S_{FG}$, which is a function of a probability distribution on a fictional ensemble. This is the cause of many other deep differences between the Gibbsian and Boltzmannian approaches. For instance, the two approaches conceive of equilibrium quite differently and, as a consequence, they also conceive of the SM counterparts of the thermodynamic laws differently. For Gibbs, equilibrium is characterized by a particular probability distribution over Γ. At equilibrium $S_{FG}$ will not fluctuate. By averaging over an ensemble, it has already built into its conception of equilibrium all of the improbable members of the ensemble whose $S_B$ spontaneously decreases. Thus, in Gibbs approach, it is correct to say one has fluctuational phenomena *at equilibrium*, so the entropy at equilibrium remains constant as in thermodynamics. On Boltzmann's approach, by contrast, one fluctuates *away from* equilibrium, that is, away from the most probable value. This is because we characterize equilibrium by a particular proper subset in $\Gamma$, the one whose points all correspond to a nearly Maxwell-Boltzmann velocity distribution. We know that $S_B$ must fluctuate away from equilibrium simply because it is a function of $X$. The quasi periodicy of the dynamics for $X$ implies that (for almost every microstate) $S_B$ cannot *always* remain in equilibrium.

---

[14] The μ-space procedure is not in general useful for nonideal gases. Unfortunately, due to the neglect of the Boltzmann approach, no work (to my knowledge) has been done on providing a substitute for nonideal gases. Nevertheless, since work began on this paper two different approaches have been suggested to me, one by Uffink and Janneke van Lith and another by Sheldon Goldstein. To object to the Boltzmann approach for this deficiency is therefore premature. It would be to mistake a lack of *practical sophistication* (due to its neglect) for an *in principle* difficulty. The difficulty is clearly not a problem of principle with the definition of $S_B$. There is a fact of the matter about which subsets of Γ are macroscopically indistinguishable. Furthermore, one will focus on macroscopic functions that are invariant under permutation of particles. Each subset of Γ will have a volume proportional to the number of permutations compatible with the macrostate, and when $N$ is large one or more of these subsets will become vastly larger than the rest.

Thermodynamics states that once an isolated system achieves equilibrium, it stays in equilibrium forever (until disturbed by external influences). Gibbsian SM agrees. Gibbsians suppose that the probability distribution characterizing equilibrium is time independent. Boltzmannian SM disagrees, however, as it abandons the idea that equilibrium is stationary in time. The Boltzmann approach balances this affront to thermodynamics by retaining the idea that equilibrium and entropy are properties of individual systems. The Gibbs approach pays for its strict agreement with the thermodynamic laws by relinquishing the idea that entropy and equilibrium are properties of individual systems. We are confronted with a difficult choice. As Sklar puts it, the choice is of "sticking with a concept most like the thermodynamic concept and exchanging the law of thermodynamics for a statistical regularity [Boltzmann] or of moving to an ensemble concept and keeping the unexceptionless law [Gibbs]" (*op. cit.*, p. 365). In the next two sections, I argue that there are some reasons for preferring the Boltzmannian single-system approach to the Gibbsian one when thinking about the reduction of thermodynamics to SM.

## III GETTING THE GIBBS ENTROPY MOVING

No sooner does one begin thinking about Gibbs fine-grained entropy than one notices the following consequence of Liouville's equation. Liouville's equation, $d\rho/dt=0$, a consequence of the Hamiltonian equations of motion, states that the swarm of representative phase points representing an ensemble move like an incompressible fluid. The Hamiltonian flow preserves the measure of any set of points if and only if their corresponding density function satisfies Liouville's equation. Liouville's equation implies, however, that the fine-grained entropy is invariant in time:

(3) $dS_{FG}(\rho)/dt=-k\int d\rho/dt\{1+\log\rho\}\,d\Gamma = 0$

$S_{FG}$ does not change, yet the thermodynamic entropy clearly does. This is the so-called "paradox" of the conservation of the fine-grained entropy.[15]

Although this "paradox" is standardly taken to present a huge difficulty for the theory, it is not immediately clear why this is so. The thermodynamic entropy is defined only at equilibrium. The entropy gradient holds only between equilibrium states. Explaining this fact does not require that we discover a function that is well defined and increasing during the interval between two equilibrium states. Finding a function defined only at equilibrium but whose value at a later equilibrium state is higher than at an earlier equilibrium state is enough. The Gibbs approach has no problem doing this. Since the macroscopic parameters change between two such states, we define a new Gibbsian ensemble for this state and this will match the thermodynamic entropy. The knee-jerk reaction to the conservation of the fine-grained entropy therefore needs re-examination. While, in the words of J. Bricmont (*op. cit.*), it may be "interesting or useful" to extend the concept of entropy to nonequilibrium situations, it is not necessary in order to account for thermodynamic irreversibility.

If there is a problem with $S_{FG}$, it is instead that it is "*not fair* to choose a new ensemble description of the system at a later time" (Sklar, p. 54; my emphasis). The ensemble at the later time should be, it seems, the Hamiltonian time development of the earlier ensemble; otherwise, the statement that Hamilton's equations govern the system is a fiction. Unfortunately, as Gibbs found, no interesting function of the earlier ensemble approaches the value of that function for the later ensemble. The time development of the earlier ensemble will not lead it to approach the later one in any way. This fact is a more serious worry about the fine-grained entropy, not that it does not monotonically increase in the intervals between equilibrium states. The problem, therefore, really has much more to do with whether the fine-grained entropy can provide a plausible reduction basis for the thermodynamic entropy than it does with the simple fact that $S_{FG}$ does not move.

In any case, my present target is not the fine-grained entropy itself, but rather the multitude of attempts to get $S_{FG}$ moving between equilibrium states. Beginning when the Ehrenfests first complained about Gibbs's treatment of irreversibility, the last eighty-five years have witnessed the near universal acceptance of the idea that to explain "irreversibility," we must find some SM entropy function that behaves monotonically in time. These approaches either retain $S_{FG}$, and therefore modify the dynamics to escape Liouville's theorem, or they define a new Gibbs-like entropy and try to show that it behaves monotonically with time. The goal, which has become the Holy Grail of modern research in the foundations of SM, is to find an entropy function that strictly mirrors the thermodynamic entropy. In other words, the goal is to find some function that always increases until reaching a maximum value at which it remains forever. Despite producing many interesting results, I believe these approaches are unsuccessful, unnecessary, and wrong-headed. Showing that this is the case is the task of the remainder of this section and the next.

---

[15] A. Wehrl, "General Properties of Entropy," *Reviews of Modern Physics*, L (1978): 221-58, here p. 227.

Before getting into the arguments, let us first ask *why* one would demand having a *strict* statistical mechanical version of the Second Law, one exhibiting non-TRI behavior. Why believe nomic irreversibility is necessary for the task at hand? One often finds the claim that thermodynamic entropy increase is strictly irreversible, and therefore, that we must base any mechanical counterpart of entropy on something not TRI. This belief is based on a deep confusion about what it takes for one theory to correspond to another. All we need to do is explain how something *apparently irreversible* emerges from something reversible, and we *need not* accomplish this through non-TRI laws. We merely need to reproduce the *image* of thermodynamics in SM at the appropriate scale, not the actual mathematical properties of the thermodynamic laws. When trying to recover one higher-level theory from a lower-level theory, we merely need to recover the phenomena supporting the theory of interest, not a literal statement of the reduced theory. TRI laws plus temporally asymmetric initial conditions can of course mimic the phenomena governed by thermodynamics, so non-TRI laws are not necessary.[16]

Despite the great effort put into getting $S_{FG}$ moving, no successful approach has yet emerged, nor does any seem likely to emerge. Understanding irreversibility this particular Gibbsian way puts one in a hole out from which one cannot plausibly dig. Obviously, I cannot do justice to the rich variety of such attempts in one essay, never mind one section of an essay. Fortunately, most of the major objections to the alternatives are already well documented. For this reason, I shall quickly remind the reader of some of these difficulties, refer her to the relevant readings, and then make a more general objection. I think it is primarily important that the reader get a flavor of the *difficulty and cost* of extracting oneself from the "problem" of the conservation of $S_{FG}$.

---

[16] Some–like N. Krylov, *Works on the Foundations of Statistical Physics* (Princeton: University Press, 1979)–object that the reliance on initial conditions in the recovery of the thermodynamic laws would threaten the lawfulness of thermodynamics. If we found the laws of thermodynamics to be dependent upon special mechanical initial conditions, they would not be universally valid and would therefore lose their explanatory force (for example, their ability to support counterfactuals). This strikes me as an overreaction. If the laws of thermodynamics can only be recovered mechanically by adopting special initial conditions, then we shall merely have discovered that thermodynamics is a special science. Thermodynamics would be like biology in its need of special mechanical initial conditions. But that does not imply that thermodynamics loses its modal force any more than the same situation implies biological generalizations have no modal force. (Or if it does imply that, then the problem is already a huge one In either case, the point is that there is no *special* problem here for thermodynamics See my "What Is 'The Problem of the Direction of Time'?" *Philosophy of Science* (Supplement), LXIII, 2 (1997): 223-34.)

*Coarse-graining.* The usual response to the conservation of $S_{FG}$ is to devise new notions of entropy and equilibrium, in particular, the coarse-grained entropy and a notion of equilibrium suitable for it. The motivation for these new notions is solely as a means of escaping the above "paradox," though it is usually defended with appeals to the imprecision with which we observe systems.

In the coarse-grained approach, one divides Γ into many small finite cells of volume $\Omega_i$. One takes the average of ρ over these cells, obtaining the coarse-grained probability $\bar{\rho}$, and attributes the value of $\bar{\rho}$ to all $X \in \Omega_i$. By substituting $\bar{\rho}$ for ρ in (1), one defines the coarse-grained entropy $S_{CG}(\bar{\rho}(\rho, \Omega_i, t))$. This entropy need not be constant in time. Due to Liouville's theorem, the volume ρ remains constant; however, the density of ρ in each cell $\Omega_i$ will change. Equilibrium is now defined as the state in which ρ has fibrillated uniformly throughout the available Γ. One then hopes to demonstrate that $S_{CG}$ will monotonically approach equilibrium.

One such attempt is Tolman's generalized H-theorem. Tolman shows that, if one assumes that $\bar{\rho}_0 = \rho_0$, where $\bar{\rho}_n$ means $\bar{\rho}$ at time $t=n$, then it will follow that $S_{CG}(\bar{\rho}_0) \leq S_{CG}(\bar{\rho}_1)$, and that $S_{CG}(\bar{\rho}_0) \leq S_{CG}(\bar{\rho}_2)$, and so on. The proof is notoriously weak, however. First, it is not a consequence that $S_{CG}(\bar{\rho}_1) \leq S_{CG}(\bar{\rho}_2)$. Nothing follows about the relative values of two $S_{CG}$s if both are later than the initial condition. Second, the assumption that $\bar{\rho}_0 = \rho_0$ is a very strong and contentious one. Third, and perhaps most worrying, the irreversible behavior of $S_{CG}$ arises almost solely due to the coarse-graining. Although coarse-graining will undoubtedly be part of our explanation of entropy increase, due to the large-scale differences between micro and macro, surely it is not the whole story. Thermodynamic behavior does not depend for its existence on the precision with which we measure systems. Even if we knew the positions and momenta of all the particles in the system, gases would still diffuse through their available volumes.

Others have tried to resurrect this style of justification by using arguments that assume a Markov condition. If the transition probabilities for $\bar{\rho}$ do not depend on the actual past state of the system, then it is possible to show that this kind of rerandomization will spread $\bar{\rho}$ uniformly throughout $\Gamma$. The question is whether this sort of rerandomization posit is compatible with the underlying deterministic dynamics. To answer this, many physicists and mathematicians turn to the idea that the underlying dynamics is "mixing."

*Mixing.* In an attempt to make plausible the idea that a Markovian postulate can describe correctly the ensemble evolution, many researchers appeal to certain features of the dynamics. The property of mixing is particularly popular. Loosely put, an ensemble is mixing if, as $\bar{\rho}(t)$ spreads out in $\Gamma$, its 'threads' are uniformly distributed in the limit when $t \rightarrow \infty$. Advocates of this approach hope that real systems are mixing. If they are, this is supposed to warrant the use of a Markov postulate in the coarse-grained approach.

Despite many beautiful results, the relevance of mixing to the approach to equilibrium is doubtful. Apart from the problems of being valid only in the infinite time limit, the mixing justification also suffers from the well-known "measure zero" problem. Perhaps worse, mixing is also neither necessary nor sufficient for thermodynamic behavior. Many systems exhibiting no "irreversible" behavior have the property of mixing (for example, a particle in a rigid one-dimensional box).[17] And due to the KAM theorem, we know many real thermodynamic systems are not mixing systems (Sklar, pp. 169ff.).

*Master equations.* Master equations are a kind of generalization of Boltzmann's kinetic equation. They are linear Markovian rate equations for the probability distribution of a system. The idea here is to show that the probability distribution for a system approaches the equilibrium distribution under certain conditions. These demonstrations are successful only when one uses the infinite time limit, continually rerandomizes the system, and posits special random initial ensembles. The justification for any of these conditions and the relevance of the result if successful is obscure.[18]

*Interventionism.* No realistic thermodynamically closed system is mechanically closed. Treating them as such is clearly an idealization. A real gas in a box is subject to all sorts of ignored, uncontrollable perturbations: for example, the gravitational pull of the moons of Jupiter, quantum fluctuations of the walls of the container, cosmic rays, and so on. Interventionists take these perturbations as essential to why entropy increases. They hold that these uncontrollable perturbations warrant abandoning Hamiltonian dynamics and hence Liouville's theorem. This rejection of the classical dynamics allows them to write an equation for the fine-grained entropy which does not have the effect of leaving it an invariant of the motion. Modeling the "random" external influences with a stochastic equation of motion, perhaps as P. Bergmann and J.L. Lebowitz[19] do, they are able to circumvent the conservation of the fine-grained entropy. They therefore have no need to "brave the ambiguities of the coarse-grained entropy formulae," as Penrose puts it (*op. cit.*, p. 1938). (In terms of our earlier problem, interventionists deny assumption (C), but most deny (A), too.)

---

[17] S.K. Ma, *Statistical Mechanics* (Singapore: World Scientific, 1985), p. 446.

[18] See Sklar for critical review.

As an answer to the foundational questions in SM, however, I think interventionism is a nonstarter. If classical or quantum mechanics is the fundamental theory of the world, then it governs the environment of closed systems, too. If so, then Hamilton's and Liouville's equations govern the environment, too, and thus the total fine-grained entropy of system + environment does not increase. One might then refer to the fact that the system + environment is itself subject to uncontrollable external perturbations from its environment, and that with respect to these perturbations, the entropy of the combined system increases, and so on a la Reichenbach. But clearly this is a mug's game. Unless one is willing to bite the bullet and claim that the TRI quasi-periodic laws of mechanics only govern the universe as a whole, and not its subsystems, then this method of pushing the explanation back further and further will always be in vain. But to do this is to claim that the laws of nature do not govern the very items that are evidence for believing the laws of nature, that is, the subsystems of the world. The fundamental problem with interventionism is that it wants to have it both ways. It does not want to offer new fundamental laws (for example, as Ilya Prigogine or David Albert[20] do) but it does not wish to be constrained by the current TRI deterministic laws either.

---

[19] "New Approach to Nonequilibrium Processes," *Physical Review*, XCIX (1955): 578-87. See T.M. Ridderbos and M. Redhead, "The Spin Echo Experiments and the Second Law of Thermodynamics," *Foundations of Physics*, XXVIII (1998): 1237-70, and references therein for defenses of interventionism.

[20] Albert points out that according to the GRW interpretation of quantum mechanics, the indeterministic and non-TRI laws of nature may be helpful to an understanding of irreversibility; see "On the Foundations of Quantum Mechanics and the Approach to Thermodynamic Equilibrium," *British Journal for the Philosophy of Science*, XLV (1994). 669-77. Prigogine, by contrast, wants to change even classical mechanics; see "A Unified Foundation of Dynamics and Thermodynamics," *Chemica Scripta*, IV (1973): 5-32.

In addition, the approach founders on justifying the stochastic terms used in modeling the environment and on justifying the necessary time asymmetric effect of the environment. If the environment does not operate in a random way, interventionism will not work. We would have no guarantee that systems will tend toward equilibrium. (If we do not want such a guarantee, or we have other reasons to think systems will tend toward equilibrium, then interventionism is unnecessary.) But we cannot justify these random influences (as they often are) by appealing to our inability to control or predict the effects of the environment. Uncontrollable and unpredictable influences, we know well, are not the same as genuinely random influences. If deterministic laws govern the environment, they cannot be genuinely stochastic. A similar problem arises concerning time asymmetry. If the approach is to avoid implying that entropy increases toward the past, time asymmetric laws must govern the environment if the approach is not to beg the question. Again, unless we adopt new, in this case non-TRI, laws of nature for the environment, this assumption will not be warranted. Finally, we can ask whether it is really plausible *or necessary* to suppose that (say) the reason why gases expand through their available volumes is in part due to the locations and sizes of the moons of Jupiter.

*Radical approaches.* Given the problems with all of these approaches, it is not surprising that some physicists turn to quite radical approaches to the problem of irreversibility. One is due to Jaynes. This is the view that entropy increase is due to the informational state of the observer and the nature of statistical inference. The other is due to the so-called "Brussels School" led by Prigogine and his theory of subdynamics. This approach alters the fundamental notion of a state of a system, replacing trajectories by probability distributions. While one can sympathize with both approaches and agree that they are both ingenious in different ways, in my opinion neither view can be sustained.[22]

Finally, though each approach has its own particular difficulties to overcome, they all (with the possible exception of the Brussels school) share a very general problem, what Sklar dubs the "parity of reasoning problem." Either the reasoning behind the approaches also works for the past direction of time as well as the future, in which case the theory predicts entropy increase toward the past, or it smuggles in an unwarranted temporally asymmetric assumption, analogous to Boltzmann's notorious *Stosszahlansatz*. This assumption typically either conflicts with the Hamiltonian dynamics or merely ends up pushing the question back a step. Falling victim to the first horn of this dilemma implies the theory of interest is wrong half the time; falling victim to the second half implies the theory is either inconsistent or incomplete. For this reason, along with the multitude of less general difficulties, I consider it unlikely that any of these approaches ever will yield a satisfying explanation of irreversibility.

---

[22] For criticism of the second view, see Robert Batterman, "Randomness and Probability in Dynamical Theories: On the Proposals of the Prigogine School," *Philosophy of Science*, LVIII (1991) 241-63. For criticism of both views, see Sklar.

## IV. REDUCING THE THERMODYNAMIC ENTROPY

We have just seen how hard it is to find a Gibbsian entropy function that monotonically increases in time. But even if one manages to skirt the above objections and define an entropy with the proper behavior, that alone does not make a function suitable as a mechanical counterpart of the thermodynamic entropy; for, as Jaynes reminds us when discussing Boltzmann's H-function: "Merely to exhibit a mathematical quantity which tends to increase is not relevant to the second law unless one demonstrates that this quantity is related to the entropy as measured experimentally" (*op. cit.*, p. 392). With this sentiment in mind, I wish to make a sweeping objection against any approach that defines an entropy function that increases strictly monotonically with time.

Monotonically increasing ensemble entropies are generally regarded as the mechanical version of the thermodynamic entropy. But are they suitable for this role? Sklar notes well the difference between entropy as a feature of individual systems and entropy as a feature of probability distributions. He cautions us not to move too quickly in our reduction of the thermodynamic entropy to the ensemble entropy, remarking that this identification is fraught with difficulty. He claims, rightly I believe, that to assert an "identity" between features of ensembles and thermodynamic features would be to misrepresent the relationship between the two theories (*op. cit.*, p. 361). But he never adjudicates the debate between the single- and many- system approaches as they relate to reduction. One suspects Sklar thinks of the Gibbsian reconception of entropy as a property of ensembles rather than a property of individual systems as simply an instance of concept extension or revision. The idea of concept extension or revision is a familiar one to philosophers. When applying a concept to a new context, we may slightly reform its meaning. The reformed concept may not preserve the original term's intension or extension, yet its meaning may not have changed sufficiently to warrant calling it a new concept. What counts as sufficient? Clearly, there is no hard and fast rule telling us how to distinguish reformed concepts from new concepts; rather, we make intuitive judgments about whether the reformed concept changes from the original only in "peripheral" features. Undoubtedly, we should expect such an extension in the present case. Whenever we explain the concepts of one theory in terms of those of another we inevitably use concepts with features different from the reduced theory's concepts. When we move from thermodynamics to SM, for instance, we extend the concept of temperature to absolute negative values for spin systems. On a SM treatment of temperature, this *seems* to make sense (but see Uffink, *op. cit.*). Since "being positive" arguably is not essential to the thermodynamic temperature (like its relationship to pressure is), this appears to be a case of concept extension and not concept substitution. Yet is the switch from an entropy defined for individual systems to one defined for ensembles merely an instance of a surprising concept extension?

It may be. Perhaps we can understand features of ensembles as properties of single systems. Gibbs himself seems to have thought it possible to eliminate reference to ensembles, although it is not entirely clear what he had in mind.[23] Nevertheless, denying that entropy is a property of individual systems *in order to avoid recurrence and define a monotonically increasing entropy function* does not so much elastically stretch the original concept as it does completely break from it. We can see this in a variety of more or less striking ways. The simple argument that follows strongly suggests that no monotonically increasing function of time should be identified as the reduction basis of the thermodynamic entropy.

Suppose one is able to define a function of an ensemble, $G(\rho)$, that increases monotonically with time until reaching a maximum value, at which it remains forever. $G$ is defined for an ensemble, for we know from above that if it were not (that is, if assumption (A) were affirmed) there could be no such $G$. Finding a function like $G$, or a proof of such behavior, is the goal of many in the foundations of SM. If we could find such a $G$, we could identify it with the thermodynamic entropy, and the thermodynamic Second Law would remain a strict, unexceptional law.

Suppose we wish to describe a gas diffusing throughout its available volume into a new equilibrium state. We know that, though it does not matter to the definition of $G$, the actual gas has a microstate $X$. We also know that $X$, whatever it is, gives rise to the macrostate $M$ we see before us. These are merely the assumptions we make when we say thermodynamics is in some sense reducible to mechanics. They are completely uncontroversial. Surely, the gas has a microstate, and surely whatever microstate it occupies corresponds to the macrostate we see. Now, if we make one more meager assumption, the irrelevance of $G$ is assured. That assumption is that classical or quantum mechanics always govern the evolution of $X$. For Poincare's recurrence theorem implies that for almost all microstates, at some subsequent time in their evolution they will return either to $X$ or to an arbitrarily small neighborhood $\Delta_X$ around $X$. Since this neighborhood can be arbitrarily small, the difference between $X$ and $X + \Delta_X$ will not matter to the macroscopic description of the state. $X$ and $X + \Delta_X$ will correspond to the same M, *whatever* it is.

---

[23] See Jan von Plato, *Creating Modern Probability* (New York: Cambridge, 1994), p. 83, and references therein.

We can now exhibit the problem with the proposed identification. The behavior of our function $G$ supposedly accounts for the Second Law, even though this explanation is compatible with the real microstate of the gas being one that subsequently leads to a macrostate that the system nearly *already occupied*. $G$ will continue to increase, or remain constant, even if the direction of the thermodynamic entropy turns around. How do we know the thermodynamic entropy would turn around in such a circumstance? We know this because the thermodynamic entropy is a function of merely the endpoints of the evolution and of the macrostate of the system. Let $S_1$ be the thermodynamic entropy of macrostate $M_1$, which corresponds to the gas confined to a small section of a container. Let $S_2$ be the thermodynamic entropy of macrostate $M_2$, which corresponds to the subsequent state of the gas after it has spread evenly throughout its new volume. The thermodynamic entropy, whatever it is, is given by $\Delta S_A = S_2(M_2) - S_1(M_1)$. But when the system recurs to an earlier macrostate–in this case from $M_2$ to $M_1$–its thermodynamic entropy, whatever it is, is given by $\Delta S_B = S_1(M_1) - S_2(M_2)$. The changes $\Delta S_A$ and $\Delta S_B$ cannot be of the same sign. The evolution of the thermodynamic entropy therefore will change direction during recurrence. This is obvious, especially when we consider a film of gas diffusing run backward, but it helps to go through the reasoning to expose the uncontroversial premises of the argument. One might object and point out that the thermodynamic entropy *does not* in fact decrease. Yet that misses the point: *if* the system in question is a mechanical system it *can* recur (and *will* recur, given enough time). It is physically possible, maybe even actual, that the value of the thermodynamic entropy not even approximately match the value of $G$.

Let us now reconsider the mechanical explanation of the gas relaxing to equilibrium provided by function G. Apparently, its explanation of why the system approaches equilibrium will "work" even if the diffused gas suddenly starts to concentrate in one small corner of the container! This is the upshot of avoiding the recurrence theorem. Any theory that explains entropy increase with a function such as *G* will be forced to say its SM entropy is increasing, or remaining constant, even while its thermodynamic entropy decreases. The only other option is to give up one of the meager assumptions that created this predicament. But since these are so minimal, to abandon one of these is to give up the game.

Clearly, even without going into all the gory details concerning the many uses of the term 'reduction', we cannot say that the thermodynamic entropy reduces to *G* in any sense.[illegible] The entropy to be reduced does not supervene upon the reducing entropy, yet supervenience is usually considered a minimal condition for any reduction. Lacking supervenience, it is hard to imagine how the other metaphysical relationships philosophers sometimes desire of reduction could hold. We cannot view *G* as *causing* the thermodynamic entropy's behavior, since the two do not even covary. Nor can *G* be a posteriori *identified* with the thermodynamic entropy. There are nomologically possible worlds where the two entropies do not have the same behavior. Since any scientific identification holds across all physically (and perhaps metaphysically) possible worlds, the fact that there is a possible world where the two entropies differ implies they cannot be identical. So on any theory of reduction remotely Nagelian in spirit, where one looks for a correspondence (even approximate correspondence) between a concept at the reducing level and a concept at the reduced level, we do not have a reduction. This is the price of making *G*'s evolution independent of the individual system's evolution: *G* becomes irrelevant to the individual system. The drive to find a function that behaves like *G*, we see, runs directly contrary to the desire to have a mechanical reduction of the thermodynamic entropy.

## V BOLTZMANN REVISITED

SM need not and should not search for a mechanical counterpart for entropy that exhibits monotonic behavior with time. For when we find such an entropy, we immediately know that it is irrelevant to the behavior of real individual systems in the world. If this is right, how should we react to this conclusion?

---

[illegible] For some of the gory details, see M. Spector, *Concepts of Reduction in Physical Science* (Philadelphia: Temple, 1978) and references therein

One reaction might be to take it (either happily or sadly, depending on one's other views in philosophy of science) as announcing that we cannot have a SM counterpart of the thermodynamic entropy. The thermodynamic entropy, then, would be viewed as a concept more like 'chair' than like 'carbon'. One might then either adopt a nonreductivist or an eliminitivist stance toward the thermodynamic entropy. The options and moves available here are familiar from the philosophy of mind literature on folk-psychological terms.

I think it is a bit early, however, to give up the search for a SM counterpart of the thermodynamic entropy. The thermodynamic entropy is a real (indirectly) observable property of individual systems. That the major Gibbsian approaches to SM have not found it does not imply a mechanical counterpart does not exist. Rather, if we cannot find it in these approaches, then we ought to look elsewhere.

One entropy worthy of more investigation is the *unaltered* Gibbsian fine-grained entropy $S_{FG}$. We saw that the conservation of $S_{FG}$ is not the problem many believe it to be. Although there remain questions about it, I think it premature to close the door on $S_{FG}$ from the perspective of reduction.

Another entropy worthy of much more thought is the largely neglected Boltzmann entropy–in particular, the reconstructed, post-H-Theorem entropy recently championed by Lebowitz (*op. cit.*) and others. The philosophical literature on SM almost completely ignores this entropy. And despite a few articles articulating the theory, I believe it is fair to say that many prominent physicists remain confused about this approach. Certainly, the letters replying to the Lebowitz article make this abundantly clear.[34] The reason for the neglect lies in a complicated web of different influences. Boltzmann's later theory is often confused with his earlier, infamous H-theorem.[35] Moreover, his own writing is difficult to follow and his reasoning is often convoluted. When combined with several common philosophical mistakes regarding Zermelo and Loschmidt's objections to Boltzmann, and the generally confused state of the field, such misunderstanding is not terribly surprising. In any case, though I cannot guarantee that this approach is problem free or even well developed, I believe it represents a promising candidate for understanding why mechanical systems tend to approach thermal equilibrium. Due to its neglect and the confusion surrounding it, I believe it is worth describing the position in greater detail than I did in section II. I would like to do what I can in this section to shed some light on potentially confusing points. I shall not claim that the theory yet offers a satisfactory reduction of the thermodynamic entropy. Although I believe it to be more promising than mainstream Gibbsian approaches, I shall show that it still faces some challenges.

---

[34] *Physics Today* (November 1994): 113-15.

[35] Jaynes shows that the H-function differs (nonnegligibly) from $S_{FG}$ whenever the potential energy of a system is relevant (that is, when there are interactions). Consequently, since $S_{FG}$ agrees with $S_B$ at equilibrium, the H-function differs (nonnegligibly) from $S_B$ too.

What kind of behavior should we expect from $S_B$? Unlike $S_{FG}$, $S_B$ is *not* an invariant of the motion, despite Liouville's theorem. To see this, consider the volume of macrostate $M$ in $\Gamma$ at $T = 0$, $|\Gamma_M|$. I shall represent the time development of the points in $|\Gamma_M|$ at $T = t$ as $|\Gamma_M'|$. Liouville's theorem implies that $|\Gamma_M| = |\Gamma_M'|$. But this does not imply that the volume of the *new* macrostate $M^*$ in $\Gamma$ at $T = t$, $|\Gamma_{M^*}'|$, is equal to $|\Gamma_M'|$. Indeed, because most of the states in $|\Gamma_{M^*}'|$ will not usually have evolved from the relatively tiny region $|\Gamma_M|$, $|\Gamma_M'|$ is not only *unequal* to $|\Gamma_{M^*}'|$, but it is typically *not even close* to as large as $|\Gamma_{M^*}'|$. Thus, $S_B$ is free to vary with time.

Furthermore, it is overwhelmingly likely to increase. One can see this qualitatively as follows. Consider a gas in equilibrium confined to a small portion of a box. It will correspond to a certain volume $|\Gamma_M|$. Now, when we remove the partition holding the gas inside that small section, the values of the macroscopic variables will change, opening up a large new volume in $\Gamma$ available to the system. Since the new equilibrium distribution occupies *almost all* of the newly available phase space, almost all of the microstates originating in the smaller volume will evolve into this region $|\Gamma_{M^*}|$ of the larger phase space. Once the representative phase point of the system evolves into this larger region, it stands virtually no chance of evolving backward. Since $S_B$ is the relative measure of the amount of $\Gamma$ corresponding to a macrostate, we can see that $S_B$ will increase when the volume corresponding to $M$ increases from $|\Gamma_M|$ to $|\Gamma_{M^*}|$. Notice also that this explanation, unlike many in SM, works only for large $N$, which is reassuring. J.C. Maxwell[*] expressed the essence of this explanation when he wrote: "The Second Law has the same degree of truth as the statement that if you throw a tumblerful of water into the sea you cannot get the same tumberful of water out again" (*ibid.*, p. 583).

We need one more assumption before our sketch of Boltzmann's theory is complete; for as the reader has probably noticed, I have not yet addressed Sklar's "parity of reasoning" problem. It is still true that for typical microstates $S_B$ will increase both toward the future and toward the past from present nonequilibrium states. That is, given a nonequilibrium state, we predict that $S_B$ will increase toward the future and that the system *evolved from* a system with higher $S_B$. The solution to this problem, as Boltzmann and many other luminaries saw, lay in appealing to temporally asymmetric boundary conditions. Richard Feynman[17] notes this clearly: "it is necessary to add to the physical laws the hypothesis that in the past the universe was more ordered, in the technical sense, than it is today" (*ibid.*, p. 116). By assuming that what we call the earlier states of the universe are of comparatively low entropy with respect to what we call the later states, we remove the unwelcome retrodiction. Earlier states do not have higher entropy than present states simply because we assume that sometime shortly after the Big Bang entropy was almost unimaginably low. Following Albert, let us call this cosmological hypothesis the *past hypothesis*:

---

[*] P M. Harman, ed , *The Scientific Letters and Papers of J.C. Maxwell, Volume II: 1862-1873* (New York: Cambridge, 1995).

> Past hypothesis: the initial (or very early) macrostate of the universe is one of extraordinarily low entropy.

The past hypothesis solves the problem of the direction of time. Whether it adequately *explains* the direction of time is another question (see my *op. cit.*). It also tells us that we can live with recurrence. Given that the universe currently resides so far from equilibrium, it is overwhelmingly probable that entropy did and will increase for all relevant observation times.

Before turning to two worries about the Boltzmann approach, let me try to dispel the idea that there is anything unpleasantly subjective about $S_B$. The coarse graining used to define $S_B$ does not imply that $S_B$ is subjective, only that it is relational. The thermodynamic variables are used to characterize the macrostates of interest. Which microstates give rise to a given thermodynamic state is an objective matter. It does not rely on the precision with which we can measure a state or our knowledge of a state. If a Laplacian intelligence informed you of the exact microstate of some macrostate, that would not affect the value of $S_B$ one jot. You would have more information about the actual microstate of the system than $S_B$ could provide, but $S_B$ is completely independent of how much you know. This is true despite $S_B$'s being relational. As Jaynes (*op. cit.*) points out, a crystal of salt may have two different entropies: one expressed as a function of temperature, pressure and volume, $S(T,P)$, another expressed, say, as a function of temperature, the strain tensor and electric polarisation, $S(T,\epsilon,E)$. But this does not imply, as Jaynes thought, that entropy is anthropocentric in nature. It implies merely that Boltzmann's entropy, to its credit, reproduces an ambiguity already existing in thermodynamics. How many microstates correspond to a particular thermodynamic description is still an objective matter, even if a system admits of more than one such description.[28]

---

[17] *The Character of Physical Law* (Cambridge: MIT, 1965).

$S_B$ seems to have the credentials we want in a reduction basis for the thermodynamic entropy. It is an objective function of the state of individual systems. The actual microstate of a gas matters to the Boltzmannian explanation of thermodynamic behavior, in the sense that we can tell whether an individual system is in equilibrium or not. Further, the entropy behaves the right way. For these reasons, and some other mathematical features (for example, its additivity), $S_B$ seems very promising.[29] Before we get too excited, however, let me finish with three questions about $S_B$ which are not often discussed.

First, for the Boltzmannian approach to work the actual microstate needs to be a generic or *typical* one for $M$ (in the language of Lebowitz). We know from the argument of section I that we cannot expect entropy increase from all microstates all of the time. We also know, however, that "most" microstates will tend toward equilibrium, since the size of the regions deviating from thermal equilibrium (according to Lebesgue measure) becomes negligible for large $N$. As $N \to \infty$, the proportion of these typical states in $\Gamma$ goes to one. But since $N \neq \infty$ for real systems, the set of atypical points has positive measure. What justification do we have for ignoring microstates in this set? And are we not just begging the question if we assume the actual microstate is always in the set of typical points? Have we really accomplished anything? (This problem is similiar in some respects to the so-called "measure-zero" objection to ergodic theory—only here the atypical set is not measure zero.)

It is important not to overreact to this problem. The worry is essentially asking for an independent justification of the imposition of our "natural" probability metric on $\Gamma$, which is something we do not have. But we can only try to solve one problem at a time, and anyway, the problem of justifying the "natural" probability metric is a very large one common to all of the different approaches to SM.[30] Indeed, it is a problem wherever we use probabilities, even outside physics (for example, in games of chance). But, surely, it is *interesting and striking* to discover that according to a very *natural* measure, Lebesgue measure, *nearly all* of the initial conditions compatible with a macrostate are such that Hamiltonian evolution will take them to states closer to thermal equilibrium. And, surely, it is interesting and striking (for instance) that it can be rigorously shown that this dominant set of microstates will, in fact, have their time evolution approximated by generalizations such as the Boltzmann equation.[31] To be sure, the typicality assumption is somewhat undesirable, but given the lesson of the reversibility and recurrence objections, I do not know how we can escape ultimately making such an assumption without changing the dynamical laws.

---

[28] Another potential source of subjectivity lay in the fact that to define $S_B$ we must "discretize" $\Gamma$ into small cells So long as the number of cells grows slowly with respect to $N$, however, this will not affect the behavior of $S_B$.

[29] It is worth noting that we can extend this approach to quantum mechanics. The volume corresponding to a macrostate has a natural analogue in Hilbert space, namely, the dimension of the projector on the macrostate.

Furthermore, intuitively, the evolution of atypical microstates is very contrived, as these states systematically avoid evolving into the larger compartments of $\Gamma$. One hopes that some mild dynamical constraints that will prohibit this "contrived" behavior. Computer models and toy models such as the Kac ring model already suggest that these conditions will be weaker than ergodicity or mixing. Development of the Boltzmannian approach should include a statement of precisely what dynamical constraints one needs.

This point raises my second question: How should we understand the "accessible regions $|\Gamma_M|$" corresponding to X? After all, if our system is not ergodic (metrically transitive), then $\Gamma$ will decompose into subsets with boundaries that cannot be crossed by $X$. In what sense, then, is $|\Gamma_M|$ accessible to $X$, if $|\Gamma_M|$ overlaps one or more of these subsets? As Uffink expressed the problem to me, what dynamical relevance does the relative size of $|\Gamma_M|$ have? One partial answer might be that the compartments of $\Gamma$ representing equilibrium are *so* large that, even if $X$ is trapped in some subset of $\Gamma$, still within that subset "most" of the states are ones either representing equilibrium or states closer to equilibrium. Another hope might be that the dynamical constraints mentioned in the above paragraph will show that most of $|\Gamma_M|$ is, in fact, accessible to $X$.

Finally, and more philosophically, it is commonly thought that one theory reduces to another only if the reducing theory in some sense explains the potentially reduced theory. If we understand 'explain' as 'causally explain', for instance, can $S_B$ play a role in a causal explanation of why systems tend toward thermal equilibrium? While playing such a role may not be necessary for the success of $S_B$, certainly it would be attractive. One reason to think it cannot fill this role is that $S_B$ essentially measures the number of microstates the system is *not in* but could have been without us noticing. As this number does not say much about the nature of $X$ itself, at first glance it hardly recommends itself as the sort of thing useful to a causal explanation. But as Peter Railton[32] points out: "The stability of an outcome of a causal process in spite of significant variation in initial conditions can be informative...in the same way it is informative to learn, regarding a given causal explanation of the First World War, that a world war would have come about...even if no bomb had exploded in Sarajevo" (*ibid.*, p. 251). To be sure, it would be wrong to think that the number of states $X$ *is not* "drives" $X$ toward equilibrium. But finding out about the (typical) "inevitability" of thermodynamic behavior does carry with it modal and explanatory force. $S_B$ quantifies this modal force.

---

[30] One potential solution to this problem is found in the GRW interpretation of quantum mechanics—see Albert.

[31] See Oscar Lanford III, "The Hard-Sphere Gas in the Boltzmann-Grad Limit," *Physica* A, CVI, 70 (1981): 70-76.

CRAIG CALLENDER

London School of Economics

---

[32] "Probability, Explanation, and Information," *Synthese*, XLVIII (1981): 233-56.

## NEW BOOKS

KENNY, ANTHONY: *A Brief History of Western Philosophy.* Malden, MA: Blackwell, 1998. xiii, 365 p. Paper $24.95.

KEPNES, STEVEN, PETER OCHS, and ROBERT GIBBS: *Reasoning after Revelation: Dialogues in Postmodern Jewish Philosophy.* Boulder: Westview, 1998. ix, 163 p. Cloth $28.00.

KERNOHAN, ANDREW: *Liberalism, Equality, and Cultural Oppression.* New York: Cambridge, 1998. xiv, 130 p. Cloth $49.95, paper $14.95.

KIM, JAEGWON: *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation.* Cambridge: MIT, 1998. viii, 146 p. Cloth $25.00; *Philosophy of Mind.* Boulder: Westview, 1997. xii, 258 p. Paper $22.00.

KIROS, TEODROS: *Self-Construction and the Formation of Human Values: Truth, Language, and Desire.* Westport, CT: Greenwood, 1998. xv, 205 p. Cloth $59.95.

KOGGEL, CHRISTINE M.: *Perspectives on Equality: Constructing a Relational Theory.* Lanham, MD: Rowman and Littlefield, 1998. xiii, 311 p. Cloth $57.50, paper $22.95.

KOLM, SERGE-CHRISTOPHE: *Justice and Equity.* Cambridge: MIT, 1997. ix, 265 p. Cloth $30.00.

KRAKAUER, ERIC L.: *The Disposition of the Subject: Reading Adorno's Dialectic of Technology.* Evanston, IL: Northwestern, 1998. xvii, 211 p. Cloth $64.95, paper $24.95.

KUIPERS, RONALD A.: *Solidarity and the Stranger: Themes in the Social Philosophy of Richard Rorty.* New York: University Press of America, 1997. 137 p. Cloth $49.50, paper $26.50.

KUKLA, ANDRE: *Studies in Scientific Realism.* New York: Oxford, 1998. xi, 176 p. Cloth $45.00.

KUPPERMAN, JOEL J.: *Value...and What Follows.* New York: Oxford, 1999. vi, 168 p. Cloth $35.00.

LACHS, JOHN: *In Love with Life: Reflections on the Joy of Living and Why We Hate to Die.* Nashville: Vanderbilt, 1998. xii, 125 p. Cloth $17.95.

LAKOFF, GEORGE, and MARK JOHNSON: *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought.* New York: Basic, 1999. xiv, 624 p. Cloth $30.00.

LANCE, MARK NORRIS, and JOHN O'LEARY-HAWTHORNE: *The Grammar of Meaning: Normativity and Semantic Discourse.* New York: Cambridge, 1997. xiii, 450 p. Cloth $69.95.

LANDINI, GREGORY: *Russell's Hidden Substitutional Theory.* New York: Oxford, 1998. xi, 337 p. Cloth $85.00.

LANE, M.S.: *Method and Politics in Plato's* Statesman. New York: Cambridge, 1998. xiv, 229 p. Cloth $59.95.

LANG, HELEN S.: *The Order of Nature in Aristotle's* Physics*: Place and the Elements.* New York: Cambridge, 1998. xii, 324 p. Cloth $64.95.

LANGTON, RAE: *Kantian Humility: Our Ignorance of Things in Themselves.* New York: Oxford, 1998. xiv, 232 p. Cloth $65.00.

LARA, MARIA PIA: *Moral Textures.* Berkeley: California UP, 1998. ix, 229 p. Cloth $45.00, paper $19.95.

LARVOR, BRENDAN: *Lakatos: An Introduction.* New York: Routledge, 1998. viii, 128 p. Cloth $65.00, paper $19.99.

LEAR, JONATHAN: *Open Minded: Working Out the Logic of the Soul.* Cambridge: Harvard, 1998. i, 345 p. Cloth $35.00.

LEMONCHECK, LINDA: *Loose Women, Lecherous Men: A Feminist Philosophy of Sex.* New York: Oxford, 1997. vii, 310 p. Cloth $45.00.

LESSNOFF, MICHAEL H.: *Political Philosophers of the Twentieth Century.* Malden, MA: Blackwell, 1999. vi, 304 p. Cloth $59.95, paper $29.95.

LEVINE, ANDREW: *Rethinking Liberal Equality. From a "Utopian" Point of View.* Ithaca: Cornell, 1998. x, 140 p. Cloth $32.50.

LEVINE, PETER: *Living without Philosophy: On Narrative, Rhetoric, and Morality.* Albany: SUNY, 1998. xi, 292 p. Paper $22.95.

LEVINSON, JERROLD: *Music in the Moment.* Ithaca: Cornell, 1998. xii, 184 p. Cloth $25.00.

LEWIS, DAVID: *Papers in Philosophical Logic.* New York: Cambridge, 1998. vi, 234 p. Cloth $54.95, paper $17.95; *Papers in Metaphysics and Epistemology.* New York: Cambridge, 1999. ix, 453 p. Cloth $64.95, paper $19.95.

LINGIS, ALPHONSO: *The Imperative.* Bloomington: Indiana UP, 1998. 234 p. Cloth $35.00, paper $15.95.

LITOWITZ, DOUGLAS: *Postmodern Philosophy and Law.* Lawrence: Kansas UP, 1997. 230 p. Cloth $35.00.

LITTLE, DANIEL: *Microfoundations, Method, and Causation.* New Brunswick, NJ: Transaction, 1998. xiii, 272 p. Cloth $34.95.

LIVINGSTON, DONALD W.: *Philosophical Melancholy and Delirium: Hume's Pathology of Philosophy.* Chicago: University Press, 1998. xviii, 434 p. Cloth $68.00, paper $25.00.

LOUX, MICHAEL J.: *Metaphysics: A Contemporary Introduction.* New York: Routledge, 1998. xi, 240 p. Cloth $65.00, paper $18.99.

LUNTLEY, MICHAEL: *Contemporary Philosophy of Thought: Truth, World, Content.* Malden, MA: Blackwell, 1999. xii, 398 p. Paper $29.95.

LUTZ, MARK J.: *Socrates' Education to Virtue: Learning the Love of the Noble.* Albany: SUNY, 1998. x, 214 p. Paper $19.95.

LYNCH, MICHAEL P.: *Truth in Context: An Essay on Pluralism and Objectivity.* Cambridge: MIT, 1998. x, 184 p. Cloth $30.00.

MACK, ARIEN, and IRVIN ROCK: *Inattentional Blindness.* Cambridge: MIT, 1998. xiv, 273 p. Cloth $37.50.

MACLEOD, COLIN M.: *Liberalism, Justice, and Markets: A Critique of Liberal Equality.* New York: Oxford, 1998. 232 p. Cloth $55.00.

MANCOSU, PAOLO: *From Brouwer to Hilbert.* New York: Oxford, 1998. x, 337 p. Cloth $65.00.

MANDLER, GEORGE: *Human Nature Explored.* New York: Oxford, 1997. xi, 208 p. Cloth $29.95.

MARKOVITS, RICHARD S.: *Matters of Principle: Legitimate Legal Argument and Constitutional Interpretation.* New York: University Press, 1998. x, 456 p.

Cloth $50.00.

MARSH, JAMES L.: *Process, Praxis, and Transcendence.* Albany: SUNY, 1999. xiii, 370 p. Paper $24.95.

MARSHALL, JOHN: *Descartes's Moral Theory.* Ithaca: Cornell, 1998. xi, 177 p. Cloth $35.00.

MARTIN, ERIC, and DANIEL N. OSHERSON: *Elements of Scientific Enquiry.* Cambridge: MIT, 1998. xii, 270 p. Cloth $30.00.

MARTIN, RAYMOND: *Self-Concern: An Experiential Approach to What Matters in Survival.* New York: Cambridge, 1998. xii, 169 p. Cloth $49.95.

MARTIN, WAYNE M.: *Idealism and Objectivity: Understanding Fichte's Jena Project.* Stanford: University Press, 1998. xvii, 175 p. Cloth $45.00.

MARTINICH, A.P.: *Hobbes: A Biography.* New York: Cambridge, 1998. xvi, 384 p. Cloth $34.95.

MATRAVERS, DEREK: *Art and Emotions.* New York: Oxford, 1998. viii, 236 p. Cloth $65.00.

MAY, LARRY: *Masculinity and Morality.* Ithaca: Cornell, 1998. x, 188. Cloth $39.95, paper $15.95.

MAZLISH, BRUCE: *The Uncertain Sciences.* New Haven: Yale, 1998. viii, 328 p. Cloth $35.00.

MAZZOTTA, GIUSEPPE: *The New Map of the World: The Poetic Philosophy of Giambattista Vico.* Princeton: University Press, 1999. xviii, 267 p. Cloth $45.00.

MCCANN, HUGH J.: *The Works of Agency: On Human Action, Will, and Freedom.* Ithaca: Cornell, 1998. x, 238 p. Cloth $49.95, paper $19.95.

MCCULLAGH, C. BEHAN: *The Truth of History.* New York: Routledge, 1997. vi, 327 p. Cloth $75.00, paper $24.99.

MCGARRY, HOWARD: *Race and Social Justice.* Malden, MA: Blackwell, 1999. xi, 237 p. Paper $24.95.

MENN, STEPHEN: *Descartes and Augustine.* New York: Cambridge, 1997. xvi, 415 p. Cloth $59.95.

MEŠTROVIĆ, STJEPAN G.: *Anthony Giddens: The Last Modernist.* New York: Routledge, 1998. x, 242 p. Cloth $75.00, paper $24.99.

MIKLOWITZ, PAUL S.: *Metaphysics to Metafictions: Hegel, Nietzsche, and the End of Philosophy.* Albany: SUNY, 1998. xxv, 221 p. Paper $18.95.

MILNE, A.J.M.: *Ethical Frontiers of the State: An Essay in Political Philosophy.* New York: St. Martin's, 1998. x, 338 p. Cloth $65.00.

MOLTMANN, FRIEDERIKE: *Parts and Wholes in Semantics.* New York: Oxford, 1997. xi, 254 p. Cloth $39.95.

MORRIS, CHRISTOPHER W.: *An Essay on the Modern State.* New York: Cambridge, 1998. x, 305 p. Cloth $54.95.

NEHAMAS, ALEXANDER: *The Art of Living: Socratic Reflections from Plato to Foucault.* Berkeley: California UP, 1998. xi, 283 p. Cloth $29.95; *Virtues of Authenticity: Essays on Plato and Socrates.* Princeton: University Press, 1999. xxvi, 372 p. Cloth $65.00, paper $21.95.

NICHOLSEN, SHIERRY WEBER: *Exact Imagination, Late Work: On Adorno's Aesthetics.* Cambridge: MIT, 1997. 270 p. Cloth $27.50.

NISSEN, LOWELL: *Teleological Language in the Life Sciences.* Lanham, MD: Rowman and Littlefield, 1997. x, 258 p. Cloth $27.50.

NUCKOLLS, CHARLES L.: *Culture: A Problem That Cannot Be Solved.* Madison: Wisconsin UP, 1998. xxvi, 301 p. Cloth $55.00, paper $22.95.

OCHS, PETER: *Peirce, Pragmatism, and the Logic of Scripture.* New York: Cambridge, 1998. ix, 361 p. Cloth $64.95.

O'CONNOR, DAVID: *God and Inscrutable Evil: In Defense of Theism and Atheism.* Lanham, MD: Rowman and Littlefield, 1998. xiii, 273 p. Cloth $63.00, paper $23.95.

O'HEAR, ANTHONY: *Beyond Evolution: Human Nature and the Limits of Evolutionary Explanation.* New York: Oxford, 1997. viii, 220 p. Cloth $29.95.

OLAFSON, FREDERICK A.: *Heidegger and the Ground of Ethics: A Study of Mitsein.* New York: Cambridge, 1998. ix, 108 p. Cloth $49.95, paper $14.95.

ONUF, NICHOLAS GREENWOOD: *The Republican Legacy in International Thought.* New York: Cambridge, 1998. xiv, 284 p. Cloth $64.95, paper $24.95.

PALMER, CLARE: *Environmental Ethics and Process Thinking.* New York: Oxford, 1998. 256 p. Cloth $65.00.

PARKER, KELLY A.: *The Continuity of Peirce's Thought.* Nashville: Vanderbilt, 1998. xvi, 268 p. Cloth $39.95.

PATERSON, R.W.K.: *The New Patricians: An Essay on Values and Consciousness.* New York: St. Martin's, 1998. 168 p. Cloth $55.00.

PETRINOVICH, LEWIS: *Darwinian Dominion: Animal Welfare and Human Interests.* Cambridge: MIT, 1999. ix, 431 p. Cloth $45.00.

PHILIPSE, HERMAN: *Heidegger's Philosophy of Being: A Critical Interpretation.* Princeton: University Press, 1998. xviii, 555 p. Cloth $60.00, paper $29.95.

PITKIN, HANNA FENICHEL: *The Attack of the Blob: Hannah Arendt's Concept of The Social.* Chicago: University Press, 1998. vii, 365 p. Cloth $30.00.

POLS, EDWARD: *Mind Regained.* Ithaca: Cornell, 1998. x, 151 p. Cloth $29.95.

POLT, RICHARD: *Heidegger.* Ithaca: Cornell, 1999. xi, 197 p. Cloth $39.95, paper $18.95.

POPP, JEROME A.: *Naturalizing Philosophy of Education: John Dewey in the Postanalytic Period.* Carbondale: Southern Illinois UP, 1998. x, 153 p. Cloth $34.95.

POTTER, VINCENT G.: *Charles S. Peirce: On Norms and Ideals.* New York: Fordham, 1998. xxxi, 229 p. Cloth $30.00, paper $16.00.

QUINTON, ANTHONY: *From Wodehouse to Wittgenstein: Essays.* New York: St. Martin's, 1998. 360 p. Cloth $49.95.

RAJCHMAN, JOHN: *Constructions.* Cambridge: MIT, 1998. viii, 142 p. Paper $15.00.

RAPP, CARL: *Fleeing the Universal: The Critique of Post-Rational Criticism.* Albany: SUNY, 1998. xi, 297 p. Paper $21.95.

REDPATH, PETER A.: *Masquerade of the Dream Walkers: Prophetic Theology from the Cartesians to Hegel.* Atlanta: Rodopi, 1998. xiv, 331 p. Cloth $92.00, Paper $23.50.

RESCHER, NICHOLAS: *Complexity: A Philosophical Overview.* New Brunswick, NJ: Transaction, 1998. xviii, 219 p. Cloth $32.95; *Communicative Pragmatism and Other Philosophical Essays on Language.* Lanham, MD: Rowman and Littlefield, 1998. ix, 213 p. Cloth $65.00, paper $23.95.

REZNEK, LAWRIE: *Evil or Ill? Justifying the Insanity Defence.* New York: Routledge, 1998. xi, 329 p. Cloth $75.00, paper $22.99.

RICHARDSON, ALAN W.: *Carnap's Construction of the World: The Aufbau and the Emergence of Logical Empiricism.* New York: Cambridge, 1998. x, 242 p. Cloth $49.95.

RIDLEY, AARON: *Beginning Bioethics: A Text with Integrated Readings.* New York: St. Martin's, 1998. xv, 288 p. Paper $27.45; *Nietzsche's Conscience: Six Character Studies from the* Genealogy. Ithaca: Cornell, 1998. 163 p. Cloth $39.95, paper $16.95.

RIPSTEIN, ARTHUR: *Equality, Responsibility, and the Law.* New York: Cambridge, 1999. xii, 307 p. Cloth $54.95.

ROBERTS, HUGH: *Shelley and the Chaos of History: A New Politics of Poetry.* University Park: Penn State UP, 1997. x, 534 p. Cloth $75.00, paper $25.00.

ROBINSON, GUY: *Philosophy and Mystification: A Reflection on Nonsense and Clarity.* New York: Routledge, 1998. ix, 299 p. Cloth $75.00.

ROCKMORE, TOM: *On Heidegger's Nazism and Philosophy.* Berkeley: California UP, 1997. xvi, 382 p. Paper $19.95.

RORTY, RICHARD: *Achieving Our Country: Leftist Thought in Twentieth-Century America.* Cambridge: Harvard, 1998. x, 159 p. Cloth $18.95 ; *Truth and Progress: Philosophical Papers, Volume III.* New York: Cambridge, 1998. vii, 355 p. Cloth $59.95, paper $18.95.

ROSENFELD, MICHEL: *Just Interpretations: Law between Ethics and Politics.* Berkeley: California UP, 1998. xii, 296 p. Cloth $45.00.

ROTH, ROBERT J.: *Radical Pragmatism.* New York: Fordham, 1999. xviii, 168 p. Cloth $35.00, paper $18.00.

ROSENBERG, JAY F.: *Thinking Clearly about Death, Second Edition.* Indianapolis: Hackett, 1998. xix, 337 p. Cloth $34.95, paper $16.95.

ROSS, STEPHEN DAVID: *The Gift of Touch: Embodying the Good.* Albany: SUNY, 1998. xix, 389 p. Paper $26.95.

ROTHSTEIN, BO: *Just Institutions Matter: The Moral and Political Logic of the Universal Welfare State.* New York: Cambridge, 1998. xvi, 254 p. Cloth $59.95, paper $19.95.

ROTTSCHAEFER, WILLIAM: *The Biology and Psychology of Moral Agency.* New York: Cambridge, 1998. xi, 293 p. Cloth $59.95.

ROVANE, CAROL: *The Bounds of Agency: An Essay in Revisionary Metaphysics.* Princeton: University Press, 1998. 260 p. Cloth $37.50.

RUDMAN, STANLEY: *Concepts of Person and Christian Ethics.* New York: Cambridge, 1997. xii, 409 p. Cloth $69.95.

RUNDLE, BEDE: *Mind in Action.* New York: Oxford, 1997. xii, 301 p. Cloth

$70.00.

SALLIS, JOHN: *Shades: Of Painting at the Limit.* Bloomington: Indiana UP, 1998. xvi, 171 p. Cloth $35.00.

SALMON, WESLEY C.: *Causality and Explanation.* New York: Oxford, 1998. xiv, 434 p. Cloth $65.00, paper $29.95.

SAMAR, VINCENT J.: *Justifying Judgment: Practicing Law and Philosophy.* Lawrence: Kansas UP, 1998. xvii, 307 p. Cloth $40.00.

SARKAR, SAHOTRA: *Genetics and Reductionism.* New York: Cambridge, 1998. x, 246 p. Cloth $59.95, paper $19.95.

SASSOWER, RAPHAEL: *Technoscientific Angst: Ethics and Responsibility.* Minneapolis: Minnesota UP, 1997. xv, 140 p. Cloth $42.95, paper $16.95.

SAYRE, KENNETH M.: *Belief and Knowledge: Mapping the Cognitive Landscape.* Lanham, MD: Rowman and Littlefield, 1997. xvi, 310 p. Cloth $57.50, paper $22.95.

SCANLON, T.M.: *What We Owe to Each Other.* Cambridge: Harvard, 1998. ix, 420 p. Cloth $35.00.

SCHALL, JAMES V.: *Jacques Maritain: The Philosopher in Society.* Lanham, MD: Rowman and Littlefield, 1998. xxi, 241 p. Cloth $58.00, paper $19.95.

SCHARFSTEIN, BEN-AMI: *A Comparative History of World Philosophy: From the Upanishads to Kant.* Albany: SUNY, 1998. xiv, 685 p. Paper $29.95.

SCHMIDTZ, DAVID, and ROBERT E. GOODIN: *Social Welfare and Individual Responsibility.* New York: Cambridge, 1998. xviii, 222 p. Cloth $49.95, paper $14.95.

SCHNEEWIND, J.B.: *The Invention of Autonomy: A History of Modern Moral Philosophy.* New York: Cambridge, 1998. xxii, 624 p. Cloth $69.95, paper $24.95.

SCHOPP, ROBERT F.: *Justification Defenses and Just Convictions.* New York: Cambridge, 1998. xii, 212 p. Cloth $54.95.

SCHROEDER, JEANNE LORRAINE: *The Vestal and the Fasces: Hegel, Lacan, Property, and the Feminine.* Berkeley: California UP, 1998. xxii, 347 p. Cloth $45.00.

SCHULHOFER, STEPHEN J.: *Unwanted Sex: The Culture of Intimidation and the Failure of Law.* Cambridge: Harvard, 1998. xii, 318 p. Cloth $27.95.

SCHWARZMANTEL, JOHN: *The Age of Ideology: Political Ideologies from the American Revolution to Postmodern Times.* New York: University Press, 1998. xii, 210 p. Cloth $55.00, paper $18.50.

SCRUTON, ROGER: *The Aesthetics of Music.* New York: Oxford, 1997. xx, 530 p. Cloth $39.95.

SEARLE, JOHN R.: *Mind, Language, and Society: Philosophy in the Real World.* New York: Basic, 1998. x, 175 p. Cloth $21.00.

SEBOK, ANTHONY J.: *Legal Positivism in American Jurisprudence.* New York: Cambridge, 1998. xiv, 327 p. Cloth $59.95.

SHAVER, ROBERT: *Rational Egoism. A Selective and Critical History.* New York: Cambridge, 1999. xii, 162 p. Cloth $49.95.

SHAW, WILLIAM H.: *Contemporary Ethics: Taking Account of Utilitarianism.* Malden, MA: Blackwell, 1999. x, 311 p. Paper $26.95.

SHER, GEORGE: *Approximate Justice: Studies in Non-Ideal Theory*. Lanham, MD: Rowman and Littlefield, 1997. xiv, 185 p. Cloth $58.00, paper $22.95.

SIDGWICK, HENRY: *Practical Ethics: A Collection of Addresses and Essays.* New York: Oxford, 1998. xxii, 142 p. Cloth $35.00.

SILVER, MITCHELL: *Respecting the Wicked Child: A Philosophy of Secular Jewish Identity and Education.* Amherst: Massachusetts UP, 1998. xix, 239 p. Cloth $45.00, paper $15.95.

SIMONS, MARGARET A.: *Beauvoir and* The Second Sex: *Feminism, Race, and the Origins of Existentialism.* Lanham, MD: Rowman and Littlefield, 1999. xx, 263 p. Cloth $25.95.

SINAIKO, HERMAN L.: *Reclaiming the Canon: Essays on Philosophy, Poetry, and History.* New Haven: Yale, 1998. xiii, 338 p. Cloth $35.00.

SINGER, IRVING: *Reality Transformed: Film as Meaning and Technique.* Cambridge: MIT, 1998. xiv, 210 p. Cloth $25.00.

SKINNER, QUENTIN: *Liberty before Liberalism.* New York: Cambridge, 1998. xiv, 142 p. Cloth $34.95, paper $9.95.

SLATER, HARTLEY: *Against the Realisms of the Age.* Brookfield, VT: Ashgate, 1998. 205 p. Cloth $59.95.

SMART, NINIAN: *World Philosophies.* New York: Routledge, 1999. vii, 454 p. Cloth $40.00.

SMITH, NICHOLAS H.: *Strong Hermeneutics: Contingency and Moral Identity.* New York: Routledge, 1998. x, 197 p. Cloth $65.00, paper $22.99.

SMITH, P. CHRISTOPHER: *The Hermeneutics of Original Argument: Demonstration, Dialectic, Rhetoric.* Evanston, IL: Northwestern, 1998. xii, 364 p. Cloth $74.95, paper $24.95.

SMITH, PETER: *Explaining Chaos.* New York: Cambridge, 1998. viii, 193 p. Cloth $59.95, paper $19.95.

SMITH, QUENTIN: *Ethical and Religious Thought in Analytic Philosophy of Language.* New Haven: Yale, 1997. xii, 255 p. Cloth $35.00.

SOBEL, JORDAN HOWARD: *Puzzles for the Will.* Toronto: University Press, 1998. xiii, 212 p. Cloth $55.00.

SOBER, ELLIOTT, and DAVID SLOAN WILSON: *Unto Others: The Evolution and Psychology of Unselfish Behavior.* Cambridge: Harvard, 1998. 394 p. Cloth $29.95.

STERBA, JAMES P.: *Justice for Here and Now.* New York: Cambridge, 1998. x, 246 p. Cloth $59.95, paper $19.95.

STEINER, MARK: *The Applicability of Mathematics as a Philosophical Problem.* Cambridge: Harvard, 1998. viii, 215 p. Cloth $39.95.

STEVENSON, LESLIE, and DAVID L. HABERMAN: *Ten Theories of Human Nature.* New York: Oxford, 1998. xii, 239 p. Cloth $24.95.

STIVALE, CHARLES J.: *The Two-Fold Thought of Deleuze and Guattari: Intersections and Animations.* New York: Guilford, 1998. xxii, 361 p. Paper $19.95.

STOKES, GEOFFREY: *Popper: Philosophy, Politics, and Scientific Method.* Malden, MA: Blackwell/Polity, 1998. xiii, 197 p. Paper $21.95.

THE JOURNAL OF PHILOSOPHY

VOLUME XCVI, NO. 8, AUGUST 1999

CALCUTTA UNIVERSITY
ALIPUR
CAMPUS
3-1-2000

# ALL AT SEA IN SEMANTIC SPACE: CHURCHLAND ON MEANING SIMILARITY

It is old news that the identity conditions for contents, meanings, and the like are problematic, perhaps because there cannot be identity conditions for meanings unless there is an analytic/synthetic distinction, and there is not. We are not proposing to consider whether the notion of content identity can be made metaphysically respectable. But we want to discuss a proposal about how to save semantics if it does turn out that "same meaning" cannot be reconstructed; it is one that is currently popular in both the philosophical and the cognitive science literature.[1]

The idea is that the purposes semantic theories are supposed to serve do not actually require principles of content *individuation*; all they require is a notion of content *similarity*. On such a view, similarity of meaning is primitive relative to sameness of meaning, and the latter might be of interest only as a limiting case of the former; perhaps, indeed, a limiting case that is never achieved in practice. Meaning similarity is real and robust, there is a lot of it around, and it is metaphysically prior to synonymy. So the story goes.

We have two points to make here. First, even if a notion of content similarity were on offer, it would arguably be of no use for doing

---

[1] Texts where this sort of suggestion is endorsed are ubiquitous. For some examples: Gilbert Harman, *Thought* (Princeton: University Press, 1973); Ned Block, "Advertisement for a Semantics for Psychology," in P. French, T. Uehling, H. Wettstein, eds., *Midwest Studies in Philosophy, Volume 10: Studies in the Philosophy of Mind* (Minneapolis: Minnesota UP, 1986), pp. 615-78; S. Sloman and L.J. Rips, *Similarity and Symbols in Human Thinking* (Cambridge: MIT, 1998); Stephen Stich, "On the Ascription of Content," in A. Woodfield, ed., *Thought and Object* (New York: Oxford, 1981), pp. 153-206; Ray Jackendoff, "Conceptual Semantics," in U. Eco, M. Santambroggio, P. Violi, eds., *Meaning and Meaning Representations* (Bloomington: Indiana UP, 1988), pp. 81-97.

0022-362X/99/9608/381–403

the kinds of things that a theory of meaning ought to do. Quite generally, the kinds of explanations that semantic theories are supposed to give would not survive substituting a similarity-based notion of content for an identity-based notion. This is unaccidental; it is the result of ways in which content similarity and content identity are intrinsically different kinds of relations. Second, we shall argue that, even if a similarity-based notion of content could in principle replace an identity-based notion *salve* the explanatory power of semantic theories, that would not be much cause for rejoicing. For, as a matter of fact, nobody has the slightest idea how to construct the required notion of content similarity; and the reasons for this are much the same, mutatis mutandis, as the reasons why nobody has the slightest idea how to construct a perspicuous notion of content identity. Until recently, we would have thought this claim was untendentious. But Paul M. Churchland,[1] in a number of articles over the last decade, has proposed what he says is a schema for such a construction. That he is quite wrong to say this has been the burden of several papers of ours, to which Churchland has now likewise replied. Our view is that his reply is an egregious *ignoratio elenchi*. The second part of the present discussion is our attempt to make clear why it is.

## I. WHY CONTENT SIMILARITY IS PROBABLY OF NO USE

A question that is not asked nearly as often as it should be is this: If you had a tenable notion of similarity of meaning, what exactly would you do with it? We suspect that friends of similarity-based content more or less take for granted that, given such a notion, they could just plug it in wherever semantic theories have hitherto availed themselves of synonymy, content identity, intentional equivalence, and the like. It is (we continue to suspect) because the friends of meaning similarity assume this that they rarely bother to show how a semantics built around content similarity *could* do anything like the explanatory work that semantic theories are supposed to do.[3] In fact, a plausible prima facie case can be made against this assumption; so, anyhow, we are about to argue.

---

[1] "Some Reductive Strategies in Cognitive Neurobiology," in *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science* (Cambridge: MIT, 1989), pp. 77-110; our *Holism: A Shopper's Guide* (Cambridge: Blackwell, 1992), chapter 6; Churchland, "Fodor and Lepore: State-space Semantics and Meaning Holism," *Philosophy and Phenomenological Research*, LIII (1993): 667-72; our "Reply to Churchland," *Philosophy and Phenomenological Research*, LIII (1993): 679-83; Churchland, "Second Reply to Fodor and Lepore," in R. McCauley, ed., *The Churchlands and Their Critics* (Cambridge: Blackwell, 1996), pp. 278-83, Churchland, "Conceptual Similarity across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered," this JOURNAL, XCV, 1 (January 1998): 5-32.

We can think of four main purposes for which one might want a theory of meaning and for which such theories have historically been employed. We do not deny that there may be other uses to which they could be put; but we do think it is plausible that, if a theory of meaning failed all of the following, it would not be worth having.

1. *Satisfaction conditions.* A semantic theory for a language ought to assign satisfaction conditions to its semantically evaluable expressions; likewise, mutatis mutandis, for a theory of the intentional contents of semantically evaluable mental states.[4] Moreover, it should turn out that sentences that we are all quite certain are true *are* true under the preferred assignments of satisfaction conditions. For example, ceteris paribus there would be something badly wrong with a semantic theory that assigned truth conditions to 'Sparrows are birds', 'Water is wet', 'George Washington was president', and so on, according to which any of them expresses a falsehood.
2. *Compositionality.* A semantic theory should make clear how the satisfaction conditions for complex linguistic expressions are determined by the satisfaction conditions of their grammatical parts, thereby explaining why natural languages are typically productive and systematic. Likewise, a semantics for propositional attitudes should make clear how the satisfaction conditions for beliefs, desires, and the like are determined by those of their constituent concepts, thereby explaining why minds are typically productive and systematic.
3. *Translation.* Pretheoretic intuition has it that meaning is what good translations preserve. A semantic theory should provide a notion of meaning according to which this turns out true.
4. *Intentional explanation.* A semantic theory should reconstruct a notion of content that is adequate to the purposes of intentional (for example, belief/desire) explanation

We are about to argue that, quite likely, a similarity-based theory of meaning would not do any of these four kinds of things; not even if one assumes, qua semantic realist, that an identity-based theory of meaning would do the work if only the individuation problem for contents could be solved.

---

[3] Arguing for this assumption is a burden which semantic *eliminativists*, unlike the friends of content similarity, are of course not required to discharge. There are people around, including Churchland himself in some of his moods, who apparently cannot decide whether their metaphysical bottom line is realism about content similarity or skepticism about meaning. Our discussion is only concerned with the former doctrine.

[4] We assume that linguistic expressions and propositional attitudes are both semantically evaluable; and we are neutral as to which, if either, is ontologically prior We shall generally go back and forth between the semantics of languages and the semantics of mental states as expository convenience suggests.

*I.1. Satisfaction conditions.* Consider, to begin with, any statement (or sentence or thought) you like which you are prepared to believe is atomic, has the logical form *Fa*, and is true. Nothing much except ease of exposition turns on how notions like 'logical form' and 'atomic statement' are deployed; you choose. (In fact, it does not even matter all that much how you construe the notion 'true'; we shall return to this.) Take, for example, the sentence (S1):

S1. Nixon is dead.

By our lights,

(i) (S1) is certainly true.
(ii) 'Nixon is dead(at *t*)' is true if and only if Nixon is dead (at *t*).[5]
(iii) No semantic theory (for English) could be adequate unless it entails (ii) and is consistent with (i)

Query: What would a content-similarity based semantics say about the truth conditions of (S1)?

Well, we do not know because, of course, there are not really any such theories on offer. But we guess the story would go something like this. Think how an old-fashioned, identity-based semantic theory goes about entailing (ii). Roughly, it assigns the individual Nixon to the singular term 'Nixon'; it assigns the property of *being dead* to the expression 'is dead', and it interprets sentences of the logical form *Fa* to mean that the individual designated by '*a*' has the property expressed by '*F*'. Well, likewise, mutatis mutandis, according to the new dispensation; except that, for example, we assign to the expression 'is dead' not the property of being dead, but *a range of properties all of which are similar to being dead*, and we assign to 'Nixon' *a range of individuals all of whom are similar to Nixon*.[6]

But, clearly, to proceed in this fashion is simply to give up the hope of assigning satisfaction conditions to formulas in a way that makes, for example, sentences like (ii) come out true. For example, we suppose that being comatose on one's death bed is pretty similar to being dead; but it is not the case that, if Nixon is comatose on his death bed at *t*, then Nixon is dead at *t*. If you do not believe us, ask Doctor Kevorkian.

---

[5] From here on, we ignore issues about context sensitivity; in particular, we shall not worry about implicit indexicals.

[6] Perhaps, however, names express individual concepts, in which case 'Nixon' is assigned some property similar to the one that 'Nixon' expresses. We are not at all clear what a similarity-based semantics should say about singular terms. Since the options for predicates seem a bit clearer, from here on we shall generally run the discussion on them

We take it to be just self-evident that properties constructed out of semantic similarity will not do what notions like 'expresses', 'denotes', 'means', and so on are supposed to do in statements like 'Nixon' denotes Nixon, 'is dead' expresses the property of being dead, 'is dead' means *is dead*, and the like. The problem is simple enough: if two expressions denote, express, or mean the same, then they are coextensive; but if two expressions only denote, express, or mean *something similar*, they need not be (and typically are not) coextensive. And, it is at the heart of the standard way of assigning satisfaction conditions, that what replaces '*F*' on the left-hand side of instances of the schema '*Fa* is true if and only if *a* is *F*' must be at least coextensive with what replaces it on the right-hand side. The moral seems to be that part of the price of switching from a content-identity based semantics to a content-similarity based semantics is giving up the idea that theories of meaning specify the pretheoretically correct conditions of semantic evaluation of thoughts, sentences, and the like.[7]

A last consideration under this general head: you might suppose that you could use a similarity-based semantics to assign satisfaction conditions in something like the traditional way if you are prepared to give up an absolute notion of *truth* in favor of some correspond-

---

[7] We are assuming that the project is to replace a semantics based on intensional identity with one based on intensional similarity; but we are also assuming that the usual notion of identity of *extension* survives. That is, it is not part of the story that the revisionist semantics would appeal to some notion of similarity *of extension* where traditional semantics adverts to extensional identity. (As far as we know, none of the proposals for a similarity-based reconstruction of the notion of intension on offer in the literature does call for a similarity-based reconstruction of the notion of extension: either the issue is not discussed, or the conservative notion of extension is assumed. See, for example, Hilary Putnam, "The Meaning of 'Meaning'," in *Mind, Language, and Reality* (New York: Cambridge, 1975), pp. 215-71.)

In fact, we cannot imagine how a similarity-based notion of extension would go. The problem is that the similarity of two extensions would presumably have to be grounded not in their *degree of overlap*, but in the *similarity of the individuals* that belong to their extensions, and who knows how this notion is to be explicated? Thus, for example, the extension of 'male college senior' is, presumably, "more similar" to the extension of 'female college senior' than it is to the extension of 'rock' or of 'prime number' But none of these sets overlaps at all. By contrast, the set {male truck drivers} is less similar to the set {male college students} than it is to the set {female college students}, though the first and second probably overlap and the second and third do not.

We shall, however, briefly consider a case of a similarity-based semantic theory which proposes to dispense not only with identity of meaning but also with identity of truth value. See immediately below ingly approximate notion. On this view, for example, the right truth sentence for (S1) might be not (ii) but (iv).

> (iv) 'Nixon is dead' is *similar to being true* (that is, it is roughly true) if and only if Nixon is similar to being dead.

We do not actually know of anybody into meaning similarity who clearly endorses this, nor are we sure what, exactly, such a proposal would amount to. But whatever (iv) means, there are surely counterinstances to the schema *formulas of the form Fa are roughly true if and only if a is similar to F.* The basic problem is that 'similar to', unlike 'roughly true', is, of necessity, relativized to respects. Thus, it is true (we suppose) that dogs are similar to cats; but it is not even roughly true that dogs are cats; or that our cat is a dog,...and so on. The equivalence also fails in the other direction. It is roughly true that you cannot be the Pope if you were born in the Bronx. But being the Pope and not being born in the Bronx are not at all the same sort of thing. Both of the present authors are instances to the contrary.

*I.2. Compositionality.* Compositionality is the idea that, in the typical case, when syntactically complex expressions (/thoughts) are semantically evaluable, their satisfaction conditions are determined by the satisfaction conditions of their syntactic constituents. And it is generally agreed that the crucial condition that must be satisfied in order that compositionality should hold for a class of expressions is that the satisfaction conditions of their constituents should be *context independent.* So, the fact that 'brown dog', 'green dog', 'brown cat', and 'yellow cat' are all compositional in English is part and parcel of the facts that: 'brown' means the same in the environment '__dog' that it does in the environment '__cat'; 'dog' means the same in the environment 'brown...' that it does in the environment 'green...'; and so on. Likewise, mutatis mutandis, for 'brown Poodle dog' and the like. That compositionality requires context independence is, to repeat, the consensus view, and we simply take it for granted in what follows.

Now, the claim that the meaning of a constituent is context independent is essentially the claim that schema (S2) is valid:[4]

> S2. If *m* is part of the meaning of '*a*' and '*a*' is a constituent of '*b*', then *m* is part of the meaning of '*b*'.

---

[4] Reading 'part' as 'improper part'. So, in particular, the whole meaning of a syntactically primitive expression counts as 'part of' its meaning and as part of the meaning of any complex expression of which it is a constituent.

The point of present interest, however, is that (S3)—which is what you get if you substitute 'is similar to part of the meaning of' for 'is part of the meaning of' throughout (S2)—is surely false:

> S3. If *m* is similar to part of the meaning of '*a*' and '*a*' is a constituent of '*b*', then *m* is similar to part of the meaning of '*b*'.

So, presumably, the meaning of 'cat' is similar to part of the meaning of 'The leopard is on the mat' (cats and leopards are both felines) and part of the meaning of 'The leopard is on the mat' is similar to part of the meaning of 'The explosion caused a lot of damage' (leopards and explosions are both dangerous). But the meaning of 'cat' is not similar to any part of the meaning of 'The explosion caused a lot of damage'.

The main argument is, we think, simple and obvious: the standard explication of the compositionality of content depends crucially on assuming that the semantic property of a constituent which is preserved when it is embedded in its host is content identity, not just content similarity.

*I.3. Translation.* The relation *translates*, like the relation 'is similar to the meaning of', is plausibly intransitive. For this reason, lots of friends of content similarity think that, whatever its flaws may be, it is at least likely to be useful in the theory of translation. And, indeed, there are some who think this is the *only* work a semantic theory can legitimately be asked to do because facts about translation are the only bona fide semantic facts there are.

We think, however, that this line of thought is extremely dubious, at least for the case of infinite languages (namely, languages of infinite expressive power), which we suppose to be the only case worth caring about. Suppose that a translation theory between $L_1$ and $L_2$ is a finitely specifiable, computable function from the sentences of one to the sentences of the other, under which similarity of meaning is preserved. Then, we think there cannot be such a theory unless $L_1$ and $L_2$ are compositional. The basic idea is pretty obvious. Suppose the languages are infinite and the translation theory is finite. Then presumably the translation theory works by first listing the translation relations among the primitive bases of the two languages (it says that 'chat' means the same as 'cat' and that 'chien' means the same as 'dog', and so forth) and then providing recursive procedures for constructing the translation relations among the infinitely many complex expressions. But, we suppose, this will work only where, in each of the two languages, the semantics of the complex expressions is determined in a regular way by the semantics of the primitive basis; that is, if the languages are semantically compositional.[9] If this line of thought is right, then the arguments that hold against similarity-based compositional theories (see above) are also arguments against similarity-based translation theories.

*I.4. Intentional explanations.* Paradigm intentional explanations appeal to principles that control the relations (rational, or causal, or both) between propositional attitudes with identical or overlapping contents. Thus, the 'practical syllogism' is supposed to be captured by some such schema as (S4):

> S4. If *a*(gent) desires that *P* and believes that *not-P unless Q* (where *a* believes that *Q* is contingent on an action that he believes himself able to perform), then, ceteris paribus, *a* tries to bring it about that *Q*.

Prima facie, (S4) is sound only if the formulas that substitute for *P* are identical or synonymous throughout. (It is plausible that weaker relations like, for example, logical equivalence are insufficient even when they preserve the extensions or the relata.) Our point is that (S4) is patently *un*sound if substitutions for the schematic variables preserve only *similarity* of meaning.

Being a zerkon is similar to being a diamond; zerkons are manufactured to ensure that this is so. But Marilyn Monroe did not think that zerkons are a girl's best friend, and she was right not to think so. Likewise, Chateau Rothschild is not all *that* different from Chateau Plonk; but there are many who are prepared to pay for the one what they would not pay for the other. Likewise, you may have a strong preference as to which twin you marry, even if the similarity between the twins is notable. And so on. None of these asymmetries of preference need indicate a palate that is excessively refined. To the contrary, the mediation of responses that discriminate sharply between similar situations is a fair part of what cognition is for. A landscape without a ravening tiger may be quite similar to a landscape with one. Tigers make a living out of that; it is part of the story about why they have stripes. It may thus matter a lot which of two quite similar landscapes you are in. Your cognitive mechanisms are there in large part to help you to find out.

---

[9] But what if translation turns out to be a *syntactic* relation? That is, what if the translation of $L_1$ expressions into $L_2$ expressions is fully determined given their respective syntactic structural descriptions (including, of course, their lexical inventories; see, for example, Stephen Schiffer, *Remnants of Meaning* (Cambridge: MIT, 1987); Fodor, *A Theory of Content* (Cambridge: MIT, 1990), chapter 7; Lepore, "Conditions on Understanding Language," *Aristotelian Society Proceedings*, XCVII, I (1996): 41-60). This is perfectly all right with us, but notice that it is no help to somebody who says that translations ipso facto preserve similarity of meaning. The claim that syntax *determines* translation is not at all the same as claiming that syntax is what (good) translation *preserves;* what translation preserves, according to similarity-based semantics, is similarity of content. Well, if it is to do so in the case of translation between productive languages, infinitely many meaning similarity relations among complex expressions must all be projected from finitely many meaning similarity relations among primitive expressions. But, as we remarked in the text, it looks like meaning similarity is generally *not* preserved under the recursive processes that construct complex expressions from simpler ones; in general, the semantic property such processes can preserve is content *identity*. If this is right, a *syntactic* translation procedure could *not* be expected to preserve the *semantic* properties of expressions in the languages between which it operates if both languages are productive and the only semantic property defined for their expressions is meaning similarity.

The underlying principle in all this is entirely obvious:

> If $x$ and $y$ are identical, then they are identical tout court. But if $x$ and $y$ are (merely) similar, then there is always a *respect* in which $x$ and $y$ are similar, and it can matter to their rational or causal relations what respect this is.

Here is another way to make much the same point. Presumably, $F = G \rightarrow \Box (F = G)$. But, of course, the corresponding conditional for similarity is invalid. The result is that the counterfactuals that identity statements guarantee are not, in general, consequences of the corresponding similarity statements. We assume that '$P$ explains $Q$' is true only where 'if $P$, then $Q$' does support appropriate counterfactuals. (Likewise, mutatis mutandis, for '$P$ confirms $Q$'.) All this being so, it is hardly surprising that, in general, you cannot replace content identity with content similarity in an intentional theory *salve* explanatory power. There may be some way for a similarity-based semantics to save the kind of intentional explanations that the practical syllogism provides; but we do not ourselves know how; and we have not seen any suggestions in the literature.

We end this section by remarking that it is our guess that nothing we have said so far would move Churchland in the least. What he would reply, we continue to guess, is that we have failed to realize how very radical his revisionist proposals actually are: they are intended to abstract not only from the traditional idea that the central semantic construct is content identity, but also from the traditional inventory of problems that semantic theories are supposed to solve. Well, so be it. But we, in turn, are not moved to exchange the old way of doing semantics for a new way on the grounds that the latter, though it will not solve what used to be thought the main problems about meaning, does promise to solve some new problems that it is unfortunately not

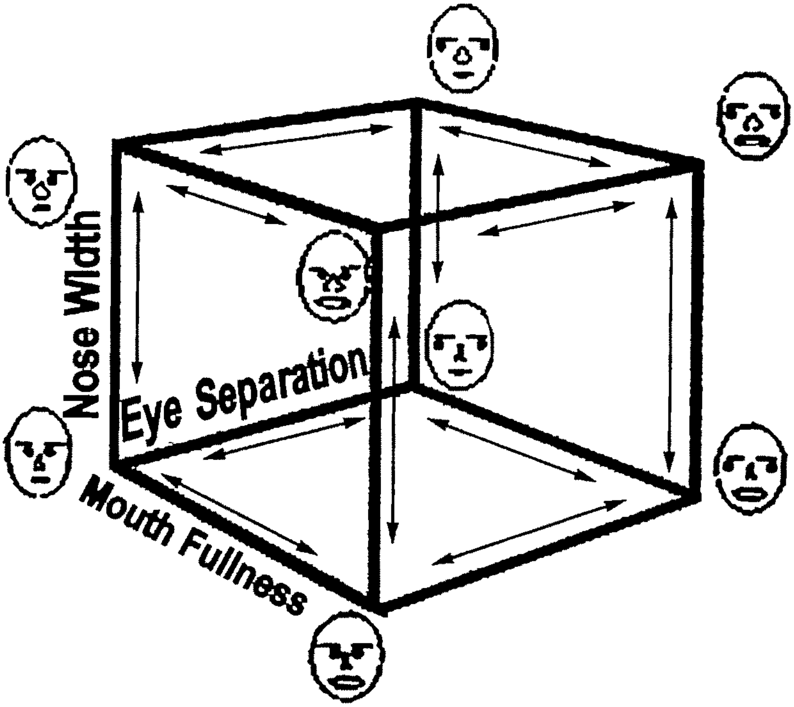


Figure 1: A rudimentary face space of three dimensions[10]

now possible to enumerate. A pig in a poke is what it sounds like to us.

## II. WHY CHURCHLAND HAS NO NOTION OF CONTENT SIMILARITY

Thus far our topic has been what role a similarity-based notion of meaning might (or might not) be able to play in a semantic theory. Suppose, for the sake of the argument, that our doubts about this are misplaced and that an account of meaning similarity would be a very good thing to have. We turn to our second topic, which is why we do not think Churchland has got one.

This requires a bit of background. Let us start with the notion of a semantic vector space. The basic idea is illustrated by Figure 1 above. This figure is supposed to provide a very simplified example of a kind of representational system that might reconstruct the perceptual similarity of faces. Roughly, there is an *n*-dimensional space in which each representation of a face occupies a proprietary region.

---

[10] After Churchland, *The Engine of Reason. The Seat of the Soul* (Cambridge: MIT, 1995), p. 28, his figure 2.5. Reproduced with permission of the author.

The dimensions correspond to properties of the faces that are salient for the perception of similarity (and/or for similarity judgments). In the present case, the space is assumed to be three dimensional for expository simplicity, "but this is highly unrealistic—our coding space for faces probably has at least twenty dimensions—but it does evoke the wide range of faces one can discriminately code with only a few resources."[11] The intended interpretation of the notation is, in effect, that locating a given face *F* at a given position *P* on a given dimension *D* corresponds exactly to representing *F* as having the property *D* to the extent *P*. The empirical claim is that, given the right choice of dimensions, the propinquity of the items in a space predicts their perceived or judged similarity.[12]

Several quick points about this kind of representation:

(1) It does not matter that, in the example, a region of the space is identified with a representation (/concept) of a *particular* face. Any representation, either of an individual or of a property, would do equally well, so long as there is a natural and "psychologically real" choice of salient dimensions along which to estimate the relative similarity of the things it represents.

(2) In fact, to claim that a concept can be properly represented as a region in a vector space really amounts to *no more* than claiming that there does exist a natural and motivated choice of such dimensions. Looked at this way, the differences between vector-space encoding and more familiar accounts of meaning is easy to exaggerate. For example, a vector-space encoding is entirely compatible with the traditional doctrine that concepts are "bundles" of semantic features. Indeed, the latter is a special case of the former, the difference being that, whereas semantic features are generally thought of as having binary values, semantic dimensions are allowed to be continuous. Unsurprisingly, in light of this, semantic feature theories yield putative metrics of content similarity just as vector-space theories do: namely, the relative similarity of two con-

---

[11] Churchland, *ibid.*, p. 29.

[12] Churchland apparently thinks that the idea of developing a geometrical system of representation, in which concepts, contents, and the like are identified with regions in a multidimensional semantic space is novel and indigenous to connectionist models of mind. "[It] contrasts sharply with the kinds of representational and processing strategies that...cognitive psychologists...have traditionally ascribed to us"—*A Neurocomputational Perspective* (Cambridge: MIT, 1991), p. 171. In fact, semantic spaces have been around for quite a long while; see, for example, C.E. Osgood, G.J. Suci, and P.H. Tannenbaum, *The Measurement of Meaning* (Urbana: Illinois UP, 1967) Most cognitive psychologists who worked with them eventually concluded that there is not any question to which they are the answer. We think they were right to conclude that.

cepts corresponds to a (possibly weighted) sum of the features they have in common.

(3) The plausibility of identifying propinquities in a space with similarities of content depends entirely on the choice of the dimensions. If, for example, you change one of the dimensions in Figure 1 above so that what it expresses is not (for example) the width of a face's nose but the weight of the face's owner, the propinquities in *that* space will not predict judgments of perceptual similarity of the faces. This is the avatar, in the geometrical notation, of the perfectly general consideration that estimates of similarity must always be relativized to "respects."

So much, then, for the system Churchland wants to employ for annotating content relations among representations. Our main claim, in earlier phases of the exchange with him (see footnote 2), was that moving to this notation requires rephrasing a number of traditional cruxes that arise for identity-based theories of meaning, but that none of them is either resolved or bypassed. We briefly recapitulate the objections we raised.

As we have just seen, similarity of content is specified only with respect to a choice of the interpretation of the dimensions of a semantic space. There are three questions to be raised about such choices, all of which are patently counterparts to classical worries about individuating content in identity-based meaning theories.

*II.1.* How is the choice of semantic dimensions to be constrained? For example, what determines that we are to taxonomize faces by the width of the noses they contain, rather than (for example) noses by some relation among the faces that contain them? Or, more plausibly, why is it right (if it is) to taxonomize the apple prototype by reference to its location in a space of colors (prototypical apples are red) rather than the red prototype by reference to its location in a space of fruits (paradigm red is apple-red). This is exactly parallel to the problem that arises in identity-based meaning theories about what constrains the choice of primitive features, as in: Why is it right (if it is) to define 'mother of' in terms of 'parent of' instead of the other way around?

Notice that this question is just as urgent for vector theories as it is for feature theories. For neither offers an interpretation of similarity relations, or indeed, of *any* semantic relations, *among* the primitive features/dimensions. Rather, they purport to capture relations of similarity of concepts *relative to* the primitive features/dimensions (that is, similarity in respect of the properties that the primitive features/dimensions express). Hence, ceteris paribus, the bigger the inventory of primitive features/dimensions they posit, the more semantic facts they fail to represent. In the limiting case, *every* concept corresponds to a primitive feature (that is, to a binary valued dimension); and the theory becomes, in effect, disquotational. This sort of point has not, of course, gone unremarked in discussions of theories of meaning. The old-speak way to put it was: the notion of a semantic feature analysis of the lexicon is trivialized if there is allowed to be a distinct feature for each lexical item (so that the semantic analysis for the concept DOG says only that it is the unique concept that has the value *plus* on the *dog* feature). Nothing about this changes when you go over from binary features to continuous ones.

We suggested, earlier in the exchange with Churchland,[13] that the reason this sort of problem seems not to bother him is that, like so many cognitive scientists, he believes deep down that only sensory properties can be semantic dimensions. This does solve the problem, but at the cost of a recidivist and deeply implausible empiricism. If Churchland has now got a better way out, we cannot find it in his text.

*II.2.* What are the truth makers for claims about the semantic interpretation of the dimensions? A second problem we raised for Churchland (in "Reply to Churchland") was that the only metaphysics of meaning he seemed to have on offer was one according to which the inferential roles of representations are the truth makers for claims about their semantic properties; and that could not be the right story about the semantic properties of the dimensions themselves.

As we saw above, to locate a concept (or a prototype) *C* with respect to a dimension *D* of a space is, in effect, to claim that the inference from being *C* to being *D* is constitutive (statistically or definitionally) of *C*. But if you thus identify the content of *C* by reference to its inferential relation to *D*, you cannot also identify the content of *D* by reference to its inferential relation to *C*. (You could, to be sure, identify the content of *D* with its position in some *higher-order* semantic space; but that would initiate a regress.) This is, once again, exactly parallel to a classic worry about identity-based meaning theories: if the contents of complex concepts are identified by their inferential relations to primitive concepts, the content of the primitive concepts has to be identified in *some other way;* and nobody knows what other way that is.

As far as we can tell, Churchland is now clear that the notion that contents are somehow individuated by their inferential roles will not work for the content of the dimensions. Rather, he thinks, the semantics of the dimensions is somehow to be reconstructed in terms of causal relations between, on the one hand, the representations that have values on the dimensions and, on the other, facts about the world. We have no objection to his taking this way out; but he does now need some reason for not treating such causal relations as the truth makers for *all* claims about the semantic properties of representations, thereby giving up, inter alia, the identification of contents with positions in semantic spaces. If (for example) the *redness* dimension has the content it does in virtue of causal relations between red things and the mental representation RED, why should the mental representation DOG not have the content it does in virtue of its causal relation to dogs? Note the exact correspondence to a traditional question about identity-based meaning theories of the 'two factor' variety: If it is to be assumed that a one-factor, causal theory works for the semantics of the primitive expressions, why is it to be denied that a one-factor causal theory will work for the semantics of *all* the expressions?

---

[13] *Holism: A Shopper's Guide.*

*II.3.* Strictly speaking, the vector theory identifies similarity of content only with respect to concepts that have rankings on the same dimensions. But it is implausible that this is true of the concepts people actually entertain. This raises a problem for similarity-based meaning theories which is exactly parallel to the traditional problem identity-based theories have about how to avoid relativizing the individuation of concepts to the belief systems in which the concepts are embedded.[14]

Suppose, for example, your concept NEBRASKA is just like mine, except in the following respect: you think that Nebraska is about a thousand miles from Guadalajara; whereas I, never having heard of Guadalajara, have no views about the distance between Nebraska and it. So, the situation is that your semantic space, unlike mine, has a dimension *proximity to Guadalajara* along which the concept NEBRASKA has a location. Question: How much (if any) of this kind of thing can be tolerated, consonant with our concepts of NEBRASKA being similar? We emphasize that the vector notation as such offers no answer to this question. Transdimensional *similarity* (similarity among concepts in spaces which differ in dimensionality) is undefined; just exactly as (lacking an analytic/synthetic distinction) an inferential-role semantics offers no interpretation of concept *identity* for minds that do not share all of their beliefs.

---

[14] Maybe you do not mind if concepts are relativized to the belief systems in which they are embedded? So be it. But, of course, the semantic relativization problem lives just across the street from the semantic holism problem: If you have to relativize the identification of concepts to minds that have *some* of their beliefs in common, how do you avoid having to relativize the identification of concepts to minds that have *all* of their beliefs in common? Our point is that, if this worry is real, it persists unaltered when you change from a traditional semantics to a geometrical semantics.

Likewise, some dimensions presumably count more than others do in similarity evaluations, and the relative weight of the dimensions differs for different concepts. So, at least in principle, it is possible for minds to differ either in respect to their weightings or in respect of the interaction between their weightings and the concepts, or both. Although you and I agree that literacy is a salient dimension for evaluating philosophers and engineers, I think it is important for philosophers to be well read, but it does not matter much whether engineers are; whereas you think it goes the other way around. Is there a principled answer to the question how (dis)similar this makes our concepts of engineers and philosophers? If so, what is the relevant principle?

There is, in short, a plethora of problems about how the similarity of concepts is to be evaluated across semantic spaces whose dimensionalities differ in various ways. These are just the old problems about the relativization of concepts to belief systems; except that, in the new notation, talk about the way that a concept is positioned with respect to the dimensions of a semantic space replaces talk of the way that a concept is embedded in a system of beliefs. Whichever notation you prefer, you have the same problems about how to identify concepts and the like in the sorts of circumstances that typically hold *across minds* (or 'languages', or 'theories').

We suspect that such problems are unsolvable unless the analytic/synthetic problem can be solved. Nonetheless, as far as we can tell, this is the issue to which Churchland takes his new theory to be germane. We are going to argue that this is a massive *ignoratio elenchi* on Churchland's part. The proposal he has on offer is not *of the right logical type* to be a solution to the problem of transdimensional content individuation; and, though there is a problem in the general vicinity that it *is* of the right logical type to solve, it does not, in fact, solve that problem either.

Churchland's proposal co-opts work by Aarre Laakso and Gary Cottrell,[13] which, he says, "constitute[s] a decisive answer to Fodor and Lepore's challenge.... [Laakso and Cottrell] successfully deploy one member of a large family of mathematical measures of conceptual similarity, measures that see past differences—even extensive differences...,"[16] and now, given the nature of the challenge that we had offered (namely, that there is a relativization problem for vector-space semantics just like the relativization problem for identity-based semantics) one expects this sentence to end "of the *semantic* dimensions of the conceptual spaces." For, to repeat, a vector-space semantics identifies concepts by reference to their positions in a semantic space; and the problem we raised—the one that Churchland apparently thinks Laakso and Cottrell have solved—was precisely 'How can minds that do not have the same semantic spaces have the

---

[13] "Qualia and Cluster Analysis: Assessing Representational Similarity between Neural Systems" (unpublished manuscript, 1998).

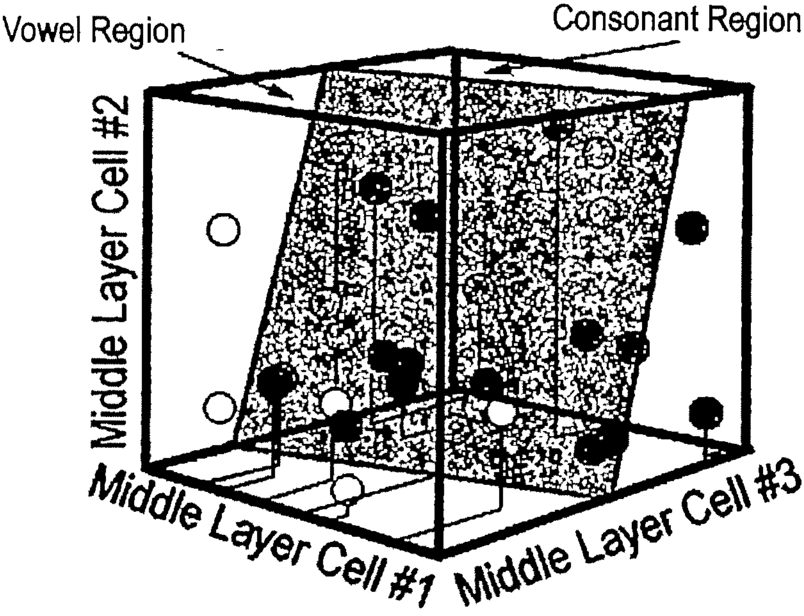


Figure 2. The activation space of the cells at the middle layer of NETtalk.[17]

---

[16] Paul Churchland and Patricia S. Churchland, *On the Contrary. Critical Essays, 1987-1997* (Cambridge: MIT, 1998), p. 81; and "Conceptual Similarity across Sensory and Neural Diversity," p. 5.

[17] From Churchland, *The Engine of Reason*, p. 88, his Figure 4.22. Reproduced with permission of the author.

same (or similar) concepts?' Here, however, is the way the sentence that we just quoted actually does end: "measures that see past differences...in the connectivity, the sensory inputs and the *neural* [our emphasis] dimensionality of the networks being compared" (*ibid.*).

Now, something has gone wrong here. Neural spaces and semantics spaces are quite different sorts of things. As we have been seeing, positions along the dimensions of the semantic space relative to which a concept is situated correspond to propositional attitudes the holding of which is supposed to be constitutive of having the concept; whereas the dimensions in a neural space *correspond to neurons*; and the positions along a given dimension specify the degree of excitation of the corresponding neuron (at a time).[14] Unsurprisingly, though he is not reliable in observing it, this difference does show up in Churchland's notation now and then; see Figure 2. Notice that the dimensions of *this* cube are labeled with the names of neurons, not with the names of semantic dimensions (compare Figure 1 above).

Now, just as there is a relativization problem about the identity (or similarity) of *concepts* and the like across *semantic* spaces of different dimensionality, so, too, there is a relativization problem about the identity (or similarity) of *brain states* across *neural* spaces of different dimensionality. For example, if you are a connectionist, you identify circuits with *n*-tuples of nodes and links; and the nodes and links in a network are themselves identified *by the totality of their relations to one another*. On this picture, the notion "same (/similar) circuit" is defined *only* for networks of identical dimensionality, just like the notion "same (/similar) content."

We think that this analogy between individuation problems has caused Churchland to confuse the question of how to provide a transdimensional notion of brain-state identity with the question of how to provide a transdimensional notion of content identity: it is because he has failed to observe this distinction that he seems indifferent whether the dimensions of the spaces he talks about are semantic or neural. It is likewise because of this confusion, and because Churchland thinks (wrongly; see below) that Laakso and Cottrell have solved the transdimensional individuation problem for brain states that he believes (again wrongly) that Laakso and Cottrell have solved the transdimensional individuation problem for content states. So we shall now argue.

---

[14] If we simplify (as, indeed, connectionist models often do) by assuming binary dimensions as opposed to ones with real number values, a state description relative to a neural space is an assignment of 1 or 0 to each neuron depending on whether it is active (at *t*)

Let us do this slowly. There is, to repeat, a problem about type individuating brain states across neural spaces of different dimensionality,[19] just as there is a problem about individuating contents across semantic spaces of different dimensionality. But it cannot simply be taken for granted that a proposed solution for the first of these problems would solve (or even help to solve) the second. Suppose you have some scheme for picking out homologous brain regions across neural spaces of different dimensionality and that you are convinced that it preserves similarity (or identity) of neural types. That is, homology is defined so that if there is a region of an $n$-dimensional space such that:

(1) All the brain states in this region are neurologically similar or identical; and if
(2) for every space of $m$-dimensions ($n \subset m$) there is a homologous region such that, if a brain state has a location in the $n$ space it has a location in the homologous region of the $m$ space; then:
(3) the brain states in the homologous regions of $m$-space are likewise neurologically similar or identical.

That is, of course, a very strong condition on a notion of transdimensional homology for regions of neurological spaces. But our point is that, even if you had a notion of transdimensional homology which, in this sense, preserved *neural* state similarity (identity), you would have no right to assume that it likewise preserves *semantic* similarity (or identity). If you want to claim that it does, you will need to provide an argument. Roughly, you need to show why, if the brain states in a region of a certain neurological space are similar or identical *in content*, then so, too, are the brain states in a region of any higher-dimension neurological space that is homologous under the preferred trandimensional mapping.[20]

---

[19] For that matter, there is a problem about typing brain states across neural spaces with the *same* dimensionality if the dimensions of the spaces are supposed to correspond to neurons as *anatomically* individuated. There really is, we think, a dilemma hereabouts. If a neuron is a *physiological* unit (for example, neurons are individuated by their connectivity and the like), then there is a holism problem for them just as there is for blatantly functional entities like circuits. On the other hand, if neurons are *anatomical* units, there is no guarantee that brains that have the same neural populations are ever in the same brain states by any functional criterion. As far as we can tell (for example, from the labeling of Figure 2 (his Figure 4.22)) Churchland prefers the first horn of this dilemma; that is, he (implicitly) takes the individuation of neurons to be sensitive to their functions. This raises, but does not solve, the question how exactly the dimension labels of his neural activation spaces are to be interpreted; that is, *which* functional properties 'being neuron $A$' and the like express.

Begging the question whether, or to what extent, the 'neurons' in their networks are supposed to be *anatomical* as well as *functional* units is a standard practice in connectionist texts. It makes it seem to uninitiates that connectionism is brain science, not just computational psychology. That is very good for getting grants, but it does cause a lot of confusion.

Notice that taking physicalism for granted (which, for the sake of the argument, we are happy to do) does *not* provide the argument that is required. "Token" physicalism says that, if individual brain states are identical, then so, too, are the corresponding individual intentional states; but token physicalism says nothing about whether there are such things as transdimensional counterparts of brain-state tokens; or whether, if there are, they ipso facto share any of their semantic properties; or, if they do, which ones. Likewise, "type" physicalism guarantees that, if brain states are of the very same kind, then so, too, are the corresponding content states. But it makes no commitments about what, if any, content relations obtain for brain states that are merely *transdimensionally homologous* according to some or other typological scheme. Churchland must be assuming a typology of brain states that meets the condition that when neurological similarity (or identity) is preserved under a transdimensional mapping, content identity is preserved, too. Nothing else would warrant his persistent practice of putting *semantic* labels on the dimensions of *neural* spaces and vice versa. But, to repeat, that semantic similarity is preserved transdimensionally when neural similarity is preserved transdimensionally does not follow from any kind of physicalism we have heard of, nor would it seem to be remotely plausible on independent grounds.

Here, to summarize, is the dialectical situation as we understand it so far: Churchland thinks that Laakso and Cottrell have provided a notion of "kind of brain state" which preserves similarity of neural state types across spaces which differ in their neural dimensionality. And he (tacitly) infers that they have *thereby* provided a transdimensional notion of content-state similarity as well. But, to put it mildly, this inference is tendentious. Churchland needs an argument for drawing it, but provides none. We doubt that he has one that he is shy about producing. Rather, our guess is that he has simply confused *semantic* spaces (of which the typical occupants are intentional states, so described) with *neural* spaces (of which the typical occupants are brain states, so described).

---

[20] It is worth noting that the claim that content-state types correspond to brain-state types under some or other criterion of individuation of the latter is implausibly strong quite aside from issues of *transdimensional* individuation. In particular, such a claim would require that wherever states of mind overlap in their intentional contents, there is also an overlap of the corresponding brain states So, for example, there would have to be a (constituent) brain state common not just to tokens of the thought *that is red*, but also to tokens of the thought *I have decided to paint the cat red*; and to tokens of the thought *red is as far from green as blue is from yellow*, and so forth. There is, as far as anybody knows, not the slightest reason for supposing that the relation between mental contents and brain states meets this constraint even if issues of transdimensionality are ignored. In fact, as things now stand, there are *no* known examples of cases where it does. It is a nice irony that it fails notoriously to be satisfied by the relation between circuits in connectionist networks and the contents that the circuits are supposed to encode. That is exactly why connectionists have so much trouble with compositionality.

All right so far. We now propose to show that, in fact, the sort of criterion Laakso and Cottrell suggest for type individuating brain states across dimensionality differences pretty clearly does *not* preserve either identity or similarity of the contents of the states.[21] Basically, Churchland's idea is that, even if my neural space has fewer dimensions than yours, we can embed the one in the other so long as the dimensions of my space are a subset of the dimensions of yours. In which case, we can assign each of my (actual or possible) neural states to the same state type as some (actual or possible) neural state of yours. A fortiori, we can type identify (at least some) brain states across (at least some) differences among neurological spaces. That, as far as we can tell, is the burden of the following remark:

> If a pair of dimensionally diverse spaces both contain...a same-shape *n*-dimensional hypersolid..., then both spaces must each either constitute or contain an *n*-D subspace, an *n*-D hyperplane of some orientation or other, that confines each hypersolid at issue. Since those respective hyperplanes...have the same dimensionality, there will be no problem in comparing the two solids.[22]

We paraphrase (under correction, to be sure):

> *Claim Q*: If my neural space is properly embedded in yours, then (ceteris paribus) my (actual or possible) brain states are a subset of yours.

Suppose, for the moment, that *Q* is true; hence that, if a neural space of lower dimensionality is embedded in a neural space of a higher dimensionality, then each brain state that has a location in the former space has a corresponding location in the latter. Our argument will be that, even if *Q* is true, *it does not provide for the transdimensional individuation of content states.* That is because *R*—which is the analog of *Q* for *semantic* spaces—is clearly false; content identity is *not* preserved under the embedding of one semantic space in another:

---

[21] A fortiori, if the Laakso and Cottrell sort of view is right about how to individuate brain-state types, then type physicalism is false about the relation between brain-state types and intentional-state types. (Type physicalists should not be alarmed by this, however; we shall see presently that the proposal on offer for transdimensionally individuating brain-state types does not work either.)

[22] Churchland, *On the Contrary*, p. 91; and "Conceptual Similarity across Sensory and Neural Diversity," pp. 13–14.

> *Claim R*: If my semantic space is embedded in yours, then (ceteris paribus) my (actual or possible) concepts are a subset of yours.

Here is a candidate counterexample to *R* (the Nebraska/Guadalajara case discussed above is another). Suppose I have only three semantic dimensions hard-soft, black-white, heavy-light. And suppose my concept ROCK is identified with a vector that specifies a region in the space that these dimensions define. Likewise for you: your semantic space also has these dimensions (inter alia); and you too have a concept that lies within the boundaries of the space that they define. The difference between us is that, whereas these dimensions define *the whole* of my semantic space, they define only an embedded subspace of yours. Now, does it follow that our concepts ROCK are similar? We should not have thought so. For perhaps your space has a dimension ANIMACY which, by assumption, mine lacks. And suppose that you think that rocks are actually quite animate; much more animate, say, than turtles; at least as animate, indeed, as half the members of the D.A.R. Is it *still* the case that your ROCK concept is similar to mine? If so, suppose also that your space also contains a dimension for abstractness, and that you think that rocks are pretty abstract; at least as abstract as the natural numbers, say, though maybe less abstract than the real numbers. Is your concept ROCK *still* similar to mine? If there are any such cases where the right answer is "no," Churchland loses.

Our point is that, even if concepts, semantic properties, intentional states, and the like can be identified with positions in a state-space, and even if position in that state-space is preserved when the latter is embedded in a space of higher dimension, it does not follow that those (or indeed any) semantic properties, intentional states, and so on can be identified with positions in the higher-order space. That is because the conditions for the individuation of contents are nonmonotonic. This, too, is something that everybody already knew; indeed, it is just the analytic/synthetic problem in one of its familiar guises. Presumably Homer has a concept of water which was similar to ours as far as it went.[13] (Water is wet, it is potable, it freezes in the winter,...and so forth.) There remains the question whether Homer's concept of water was similar to ours tout court, and the relevant considerations are much the same ones that the corresponding question about the *identity* of our concepts would turn on: our WATER concept is embedded in a theory (a *Weltanschauung*; a form of life; blah, blah, blah) many of whose dimensions are quite different from any of the ones in Homer's semantic space. What, if anything, follows about which concepts we share with him?

---

[13] Assuming, as Churchland of course does, that what concepts you have depends on what beliefs you have.

Recapitulation: even if we did have a transdimensional notion of "brain-state similarity," it would not follow that we would then have a transdimensional notion of "content similarity." That is, even if regions of a lower dimensional space can be securely correlated with semantic interpretations, there is no reason why such a correlation should survive any proposed scheme of transdimensional mapping.

## III. CONCLUSIONS

We close by remarking that, putting questions about content to one side, Churchland's sort of story really does not work even for the transdimensional typing of brain states. *Q* is false because the identity conditions for neural-state types, like the identity conditions for content-state types, is nonmonotonic. A brain state that is identified with a position in a neural space of *m* dimensions need not be type identical with *any* region in an *n*-dimensional neural space in which the *m*-space is embedded. This is because brain-state types are *functionally* individuated; and functional individuation is itself nonmonotonic.[24]

Suppose we have a neural space of *m* dimensions; and suppose that the function of a certain activation state of these neurons is to cause a flight response. Now, embed the *m*-space in a one of higher neural dimensionality (which is to say: put the same *m*-neurons in a bigger brain, preserving their connectivity to one another). It surely does not follow that there is any activation state of these (or, indeed, of any) neurons in the larger brain whose function is to cause flight responses. Indeed, it does not even follow that activation states defined over the original *m*-neurons have *any function at all* in the larger brain; that depends, inter alia, on what *other* neurons they are connected to. It is, for example, perfectly possible there is *no* physiologically distinguishable system that all or only the original *m*-neurons belong to when they are in the bigger brain.[25]

[24] We are not, please note, asserting—or even suggesting—that content individuation is nonmonotonic *because* it is a species of functional individuation. Heaven forfend.

[25] Suppose, however, that it is denied that the individuation of kinds of brain states is functional. Even so, it is far from obvious that either identity or similarity of brain-state kinds would be preserved under the sort of transdimensional mapping Churchland imagines. It is patently not the case that brain states that are similar *in a given world* because they share certain nonfunctional properties are ipso facto similar in *any* world in which they have those properties in common. There

Brief conclusion: the question we were worried about in our earlier exchanges with Churchland was: 'How can there be *content similarities* between minds that have different beliefs?' We were worried about this because Churchland keeps saying that similarity semantics avoids the main problems with which analytic/synthetic semantics is plagued, and our understanding is that 'How can there be content identities between minds that have different beliefs?' *is* the main problem with which analytic/synthetic semantics is plagued.

But the question Churchland actually offers an answer to is: 'How can creatures be in similar brain states if they have different numbers of neurons?' This is an interesting kind of question, to be sure; but it has nothing in particular to do with meaning. To the contrary, it arises for any theory that is committed to the functional individuation of things over which it quantifies. This is so whether or not the things over which it quantifies are psychological/intentional/semantic. In particular, the question of transdimensional functional state individuation arises for any theory that is committed to kinds of states that can have multiple realizations. Compare 'How can tubes and transistors both be amplifiers?', 'How can jets and diesels both be engines?', 'How can a wood thing and a plastic thing both be pawns?', and so on.

It would be nice to have a transdimensional, similarity-preserving, type-identity criterion for brain states, and it would also be nice to have a transdimensional, similarity-preserving type-identity criterion for content states. And it would be very nice (and *very* surprising) if the two criteria turned out to be the same. But, as far as anybody knows or has any reason to believe, these two individuation problems raise quite different ontological issues, and their solutions—if, indeed, they have got solutions—are likely to be largely independent. All they have in common, as far as we can tell, is that both kinds of individuation are nonmonotonic, and that Churchland has not got a workable account of either. And nor, of course, has anybody else.

JERRY FODOR
ERNIE LEPORE

Rutgers University

---

*may* be ways of choosing canonical neural dimensions that meet this condition, but there is no a priori reason to believe that there are. To put this another way: it is one question whether brain states are individuated by their function, it is quite a different question whether, however they are individuated, they are reducible to neural states.

# ON A SUPPOSED COUNTEREXAMPLE TO MODUS PONENS*

In "A Counterexample to Modus Ponens," Vann McGee[1] contends that modus ponens is not strictly valid. As he understands the rule, it tells us that from an indicative conditional, together with its antecedent, one can infer its consequent; and it fails, he argues, when applied to a conditional whose consequent is itself a conditional. McGee presents the following example in support of this claim:

> Opinion polls taken just before the 1980 election showed the Republican Ronald Reagan decisively ahead of the Democrat Jimmy Carter, with the other Republican in the race, John Anderson, a distant third. Those apprised of the poll results believed, with good reason:
>
> (1) If a Republican wins the election, then if it's not Reagan who wins it will be Anderson.
> (2) A Republican will win the election.
>
> Yet they did not have reason to believe
>
> (3) If it's not Reagan who wins, it will be Anderson (*ibid.*, p. 462; numbering added).

He adduces two other similarly ingenious examples, having essentially the same structure as the one just cited.

Alhough McGee's examples are indeed arresting, I do not think that they establish his conclusion that modus ponens is invalid. What makes the examples plausible, I believe, is the assumption that the conditionals occurring in them satisfy the law of exportation, a principle which says that a conditional of the form $\ulcorner$If $\phi$ and $\psi$, then $\chi\urcorner$ entails one of the form $\ulcorner$If $\phi$, then if $\psi$, then $\chi\urcorner$. McGee, in fact, claims that the law of exportation, as well as its converse, the law of importation, are valid for English indicative conditionals. He contends, in addition, that the law of exportation and modus ponens cannot both hold for indicative conditionals, and he is right about this, given the manner in which he construes indicative conditionals. He is mistaken, however, in defending exportation at the expense of modus ponens: it is doubtful that there is any conditional that sustains exportation (or importation) but not modus ponens. In what follows, after showing that McGee's examples rely for their plausibil-

---

* I am grateful to Jim Brown, Jack Canfield, Elmar Kremer, Doris Olin, and David Widerker for helpful comments and suggestions.

[1] This JOURNAL, LXXXII, 9 (September 1985): 462-71.

0022-362X/99/9608/404–15

ity on the law of exportation, I argue that this law is simply not valid for conditionals of the sort McGee has in mind. I then show that, quite aside from the matter of exportation, McGee's contention that an inference like the one above counts against modus ponens hinges on untenable ideas about conditionals and their interpretation. Finally, I offer an explanation of why the law of exportation, though an intuitively plausible principle of conditional reasoning, does not hold for the relevant class of conditionals.

## I

A rule of inference is valid just in case it is truth preserving, that is, just in case it will never take us from premises that are true to a conclusion that is not. Accordingly, the inference from (1) and (2) to (3) constitutes a counterexample to modus ponens only if (1) and (2) are true while (3) is not. McGee allows that, if each of (1) and (3) is taken as a material conditional, that is, a truth-functional conditional that is true if and only if it has either a false antecedent or a true consequent, then the inference is unproblematic; for construed in that way, his example is an argument having both true premises and a true conclusion (*ibid.*, p. 464). He claims, however, that it is a mistake to suppose, in general, that an English indicative conditional has the truth conditions of a material conditional and, in particular, that (3) has them; and he adduces as evidence for the latter claim the observation that, if we were familiar with the relevant polls, we would not for a moment suppose that (3) is true, even though we might be resigned to the idea that the antecedent of (3) is false (*ibid.*, p. 464).

Doubtless, many will agree with McGee that the ordinary indicative conditional is not susceptible to a truth-functional analysis and that (3) is false despite having a false antecedent. Perhaps. It is easy to see that it would be misleading and, hence, inappropriate to assert (3) simply on the ground that its antecedent is false, but whether (3) would be false in such circumstances is another matter.[2] In my view, one can make a strong case for the idea that indicative conditionals should be taken truth functionally; but for present purposes, we may put this issue to one side. I find it doubtful that there is any consistent interpretation of the conditionals occurring in

---

[2] In his 1967 William James Lectures, H. P. Grice attempts to reconcile our use of the indicative conditional with a truth-functional analysis, using principles of conversation; the lectures have been published in his *Studies in the Way of Words* (Cambridge: Harvard, 1989).

McGee's example, truth functional or otherwise, which will make the premises true and the conclusion false.[3]

We see that the inference from (1) and (2) to (3) constitutes a counterexample to modus ponens only if (3) is a false conditional, in which case it cannot be construed as a material conditional. Let us suppose, accordingly, that the conditionals occurring in this inference are not truth functional and, moreover, that (3) is false. Clearly, (2) is true. But what about (1)? It is clear that (1) is true if it is taken as a material conditional. But what reason is there for supposing that (1) is true if it is taken as a *strong* conditional, as I shall call it, that is, as an indicative conditional taken non-truth-functionally? McGee observes that someone familiar with the relevant polls is likely to believe (1), and do so with good reason. Of course, sometimes we repudiate a premise when we realize that it leads to an utterly incredible conclusion. According to McGee, this is not the case here; he contends that even in the presence of (3), one is not inclined to relinquish (1) (*op. cit.*, p. 463). This is certainly correct, but it does not, by itself, show that the conditional we are prepared to accept, when we accept (1), is a strong conditional, rather than a material conditional. And, of course, if (1) is taken truth functionally and (3) not, then the inference from (1) and (2) to (3), though having true premises and a false conclusion, is not an instance of modus ponens: since the sentence 'If it's not Reagan who wins, it will be Anderson' does not have the same meaning in the premise in which it occurs as it does in the conclusion, it would be an equivocation to infer (3) from (1) and (2).[4]

If, as McGee contends, the law of exportation is a valid principle for strong conditionals, we have a good reason for thinking that (1) is true taken as a strong conditional. In the circumstances under consideration, we are interested in the question of how the election will turn out (as opposed to, say, how the primaries might have turned out). There are exactly two Republican candidates, Reagan and Anderson. Thus, if we assume that a Republican other than Reagan will win the election, we shall naturally conclude that Anderson will win, and rightly so. In other words, it is entirely reasonable, and perfectly correct, to believe that:

---

[3] Walter Sinnott-Armstrong, James Moor, and Robert Fogelin note that while McGee says there is no reason to believe that (3) is true, he never explicitly argues that it is false—"A Defense of Modus Ponens," this JOURNAL, LXXXIII, 5 (May 1986): 296-300. They point out that for McGee's example to be a counterexample to modus ponens, (3) must be false, adding that (3) "can be true even if there is no reason to believe it" (p. 297). (They go on to suggest that McGee's target is not modus ponens at all, but rather an epistemic or probabilistic analogue of that rule.) It is doubtful, however, that anyone would think that (3) is true unless she were to give it a truth-functional reading, and, as I indicated earlier, McGee does explicitly argue that it is implausible to construe the indicative conditional as a material conditional. However this may be, I think that modus ponens is defensible on any coherent reading of the indicative conditional.

[4] E. J Lowe makes this last point in "Not a Counterexample to *Modus Ponens*," *Analysis*, XLVII, 1 (January 1987): 44-47. Although Lowe is certainly correct that McGee's example does not impugn modus ponens if (1) is taken as a material con-

(4) If a Republican wins the election and it is not Reagan who wins, then it will be Anderson.

This last conditional is certainly true when construed as a material conditional, but it is also true construed as a strong conditional. Given the law of exportation, however, we may infer (1) from (4). Therefore, if the law of exportation holds for strong conditionals, (1) is true when taken as a strong conditional.

It looks as though two intuitively plausible principles of classical logic, modus ponens and the law of exportation, are in conflict, and a potential casualty of the clash is the strong conditional. We are assuming that the conditionals that figure in McGee's example are strong conditionals and, in addition, that (2), the second premise, is true, while (3), the conclusion, is false. We have just seen that (4), taken as a strong conditional, is true. If the law of exportation is a valid principle for strong conditionals, then the inference from (4) to (1) shows that (1) is true; but if (1) is true, McGee's example has true premises and a false conclusion, in which case modus ponens is not strictly valid. On the other hand, if modus ponens is beyond reproach, then the inference from (1) and (2) to (3) shows that (1) is false; but if (1) is false, the inference from (4) to (1) has a true premise and a false conclusion, in which case strong conditionals do not satisfy the law of exportation.[5] We must conclude that at least one of these principles fails for the strong conditional. Since both principles are valid for the material conditional, and since both are intuitively plausible, someone might read these results as an argument for abandoning the strong conditional altogether: perhaps

---

ditional, I do not see that he provides a convincing reason for thinking that it must be taken as a material conditional Lowe contends that it is more plausible to suppose that the consequent of (1) is a material conditional than it is to suppose that the inference in question invalidates a deeply rooted principle of deductive inference such as modus ponens.

[5] Doris Olin uses an example similar in structure to McGee's to draw the conclusion that exportation is not valid for subjunctive conditionals—"Newcomb's Problem. Further Investigations," *American Philosophical Quarterly*, XIII, 2 (April 1976): 129-33. She argues that while the conditional 'If *I* will live 20 years after *t* and *I* were to jump off the Empire State Building at *t* (with no protective devices), then *I* would become the object of intense medical interest' is true, the conditional 'If *I*

English indicative conditionals simply have the truth conditions of a material conditional after all. Although I would not find this conclusion disagreeable, the problem is that while it would solve our immediate difficulty concerning the likes of (1), (3), and (4), a similar one would arise in connection with subjunctive conditionals, and it is hard to imagine that we would resign ourselves to the idea that, for example, the counterfactual (subjunctive) conditional 'Had Reagan not won, Anderson would have' is true.

Evidently, if we are to preserve a strong conditional, we must relinquish one of these principles of classical logic. McGee recognizes that modus ponens and the law of exportation will not both come out valid on a nonclassical account of the conditional; he thinks, however, that we have good reasons for preferring exportation to modus ponens. He says, on the one hand, that we have "explicit examples to show that the indicative conditional does not satisfy modus ponens" (*op. cit.*, p. 466) and, on the other, that we have good inductive evidence that the law of exportation holds for indicative conditionals:

> In general, we assert, accept, or believe a conditional of the form $\ulcorner$If $\phi$, then if $\psi$, then $\theta\urcorner$ whenever we are willing to assert, accept, or believe the conditional $\ulcorner$If $\phi$ and $\psi$, then $\theta\urcorner$. It appears, from looking at examples, that the law of exportation ... is a feature of English usage (*op. cit.*, pp. 464-65).

He adds that "[i]t would appear that the law of importation, the converse of the law of exportation, is also valid" (*op. cit.*, p. 465, footnote 5). McGee goes on to claim that any theory of conditionals which denies the law of exportation, such as Robert Stalnaker's,[6] cannot be entirely correct (*op. cit.*, p. 466).

Obviously, if (1) and (2) are true and (3) is false, we have an example showing that conditionals of the relevant sort do not satisfy modus ponens. It is a mistake, however, to imagine that McGee's argument against modus ponens stands on its own, independently of

---

will live 20 years after *t*, then if *I* were to jump off the Empire State Building at *t* (with no protective devices), *I* would become the object of intense medical interest' is not; she says that "even though '*I* will live 20 years after *t*' is true, we would not want to grant 'If *I* were to jump off the Empire State Building at *t* (with no protective devices), *I* would become the object of intense medical interest'" (p. 131).

[6] Stalnaker sets out his account of conditionals in "A Theory of Conditionals," in *Studies in Logical Theory, American Philosophical Monograph Series*, Number 2 (Malden, MA: Blackwell, 1968), pp. 98-112; reprinted in W. L. Harper, Stalnaker, and G. Pearce, eds., *Ifs* (Boston: Reidel, 1981), pp. 41-55. On Stalnaker's account, a conditional $\ulcorner$If $\phi$, then $\psi\urcorner$ is true in possible world *i* just in case $\psi$ is true in the possible world most similar to *i* in which $\phi$ is true. Thus, a conditional of the form $\ulcorner$If $\phi$, then if $\psi$, then $\chi\urcorner$ is true in *i* just in case its consequent $\ulcorner$If $\psi$, then $\chi\urcorner$ is true in the

exportation. We have little reason to accept (1) as a true strong conditional, unless we are prepared to accept the law of exportation, and if we have little reason to accept (1) as a true strong conditional, then we have little reason to accept the inference from (1) and (2) to (3) as a counterexample to modus ponens. The idea that this inference discredits modus ponens and the idea that exportation is legitimate for strong conditionals stand together, at least in this sense: unless we accept exportation, we have little reason to accept the idea that the inference counts against modus ponens. In short, the example under discussion depends on the background assumption of exportation for its plausibility.

## II

McGee may be right that we are apt to accept a conditional of the form $\ulcorner$If $\phi$, then if $\psi$, then $\theta\urcorner$ when we accept one of the form $\ulcorner$If $\phi$ and $\psi$, then $\theta\urcorner$. We are certainly inclined to accept (1), and I think that we do so because we accept (4). This evidence, however, does not warrant the conclusion that the law of exportation is valid for strong conditionals; nor is that conclusion true.

One reason for doubting that the linguistic evidence cited by McGee shows that the law of exportation holds for strong conditionals is based on the converse of that law, the law of importation. Certainly, if we reflect on evidence of the sort that McGee sets out, as well as on our pre-analytic intuitions, we shall find as much reason to accept importation as exportation: intuitively, it is just as plausible to infer a conditional of the form $\ulcorner$If $\phi$ and $\psi$, then $\theta\urcorner$ from one of the form $\ulcorner$If $\phi$, then if $\psi$, then $\theta\urcorner$, as the other way around. We can show, however, that law of exportation and the law of importation cannot both apply to strong conditionals; for in the presence of several standard assumptions about conditionals, the law of importation is inconsistent with the view that (1) and (2) are both true and (3) is false. One assumption is that an indicative conditional of the form $\ulcorner$If $\phi$, then $\psi\urcorner$ is logically true just in case $\phi$ logically implies $\psi$. Because every sentence logically implies itself, it follows that any sentence of the form

---

$\phi$-world (as we might call it) most similar to $i$, which is to say, on this account, just in case $\chi$ is true in the $\psi$-world most similar to $\phi$-world most similar to $i$. Consequently, the truth of a conditional of the form $\ulcorner$If $\phi$ and $\psi$, then $\chi\urcorner$ does not ensure the truth of one of the form $\ulcorner$If $\phi$, then if $\psi$, then $\chi\urcorner$: the former will be true in $i$, and the latter false in $i$, if $\chi$ is true in the $\ulcorner\phi$ and $\psi\urcorner$-world most similar to $i$ but false in the $\psi$-world most similar to the $\phi$-world most similar to $i$, which is possible if there are $\phi$-worlds (in which $\psi$ is not true) and $\psi$-worlds (in which $\phi$ is not true) more similar to $i$ than the closest $\ulcorner\phi$ and $\psi\urcorner$-world.

(5) If (if $\phi$ then $\psi$), then (if $\phi$ then $\psi$)

is logically true. But assuming the law of importation, a sentence of the form of (5) entails one of the form

(6) If (if $\phi$ then $\psi$) and $\phi$, then $\psi$

Since any sentence having the form of (5) is logically true and since (5) logically implies (6), it follows that any sentence having the form of (6) must be logically true as well. It should be clear, however, that, if the antecedent of (6) logically implies the consequent of (6), then modus ponens is a valid rule of inference. But as we saw earlier, if modus ponens is valid for strong conditionals, exportation is not. Consider, for example, the result of replacing $\phi$ in (6) with (2) and $\psi$ with (3):

(7) If (if a Republican wins the election, then if it is not Reagan who wins it will be Anderson) and a Republican wins the election, then if it is not Reagan who wins it will be Anderson.

Since (7) is logically true and since a conditional is logically true only if its antecedent logically implies its consequent, it follows that it is not possible for the antecedent of (7)—that is, the conjunction of (1) and (2)—to be true and the consequent of (7)—namely, (3)—to be false. Therefore, given that (2) is true and (3) is false, we have to abandon the idea that (1) is true. But since (4) is true, this conclusion contradicts the view that the law of exportation holds for strong conditionals. Therefore, if the law of importation holds for strong conditionals, then the law of exportation does not.

This piece of reasoning shows that one cannot consistently hold (as McGee does) both that importation is valid and that the inference from (1) and (2) to (3) is not. It also shows that, if importation is valid for strong conditionals, then exportation is not. We have not, however, demonstrated that importation is valid for strong conditionals; so this objection, while showing that McGee's account requires modification, is not necessarily fatal to exportation. But the conclusion that exportation and importation do not both hold for strong conditionals casts doubt on McGee's grounds for thinking that the law of exportation does: inasmuch as the linguistic evidence favoring importation is essentially of the same character and quality as that favoring exportation, it is hard to see how we can discount the former and not the latter.

There is, in fact, a compelling reason for thinking that the law of exportation does not hold for strong conditionals, which is that, if it does, the strong conditional is truth functional. Again, since a conditional is logically true if and only if its antecedent logically implies its consequent, it follows that any conditional of the form:

(8) If (either $\phi$ or $\psi$) and not-$\phi$, then $\psi$

is logically true. But in the presence of the law of exportation, (8) logically implies:

(9) If either $\phi$ or $\psi$, then if not-$\phi$, then $\psi$

Hence, any conditional having the form of (9) is logically true. Substituting 'It is Reagan who will win' for $\phi$ in (9) and 'It will be Anderson who wins' for $\psi$, we obtain:

(10) If either Reagan wins or Anderson wins, then if Reagan does not win, then Anderson will.

Since (10) is logically true, it follows that it is not possible for the antecedent of (10) to be true and the consequent false. But since a disjunction is true if at least one of its disjuncts is, the antecedent of (10) is true; hence, the consequent must be true as well. The consequent of (10), however, is, simply (3), which we have been assuming is false. Since (8) is unexceptionable, it seems that we must either surrender the law of exportation, which sanctioned the step from (8) to (9), or drop the critical assumption that (3) is false. In fact, the pattern of inference from (8) to (9) reveals a more general difficulty, namely, that in the presence of exportation, any conditional of the form ⌜If not-$\phi$ then $\psi$⌝ is true if either $\phi$ is true or $\psi$ is, which is to say that a conditional is true if it has a false antecedent or a true consequent. All that is required for the conclusion that ⌜If not-$\phi$ then $\psi$⌝ is fully truth functional is the observation, whose truth is uncontroversial, that the converse of (9) is logically valid as well, that is, that a sentence of the form ⌜If not-$\phi$ then $\psi$⌝ logically implies one of the form ⌜Either $\phi$ or $\psi$⌝. We may infer, as a result, that the law of exportation holds only for those conditionals having the truth conditions of the material conditional; therefore, it does not hold for strong conditionals if, as we have been assuming, they are to be distinguished from material conditionals.

Two points emerge. The first, which may be surprising given the conflict between modus ponens and exportation noted earlier, is that, if the law of exportation (or importation) holds for some class of conditionals, then so does modus ponens. And the second, which may explain why it looks as though both will not come out valid for any nonclassical conditional, is that the law of exportation simply does not hold for strong conditionals.

## III

Now that we see that the law of exportation does not hold for strong conditionals, we may return to the question of whether the inference from (1) and (2) to (3) counts against modus ponens. We agreed that (2) is true and allowed that (3) is false, and we asked whether there is any reason for supposing that (1) is true if it is taken as a strong conditional. McGee's observation that anyone familiar with the relevant polls would certainly accept (1) prompted the further question of whether there is any reason for thinking that the conditional we accept, when we accept (1), is a strong conditional. Without the law of exportation, I cannot see that there is much reason for thinking that it is. Of course, it does not follow from this that (1), taken as a strong conditional, is false. There is, however, a decisive reason for thinking that it is, at least given our other assumptions, and, hence, for thinking that the inference from (1) and (2) to (3) does not constitute a problem for modus ponens. McGee contends that it is a notoriously difficult matter to specify an account of the truth conditions of the English indicative conditional. Perhaps so; but it is easy to see that (1) cannot be true. For it is an essential feature of any conditional, indicative or otherwise, that it is false if it has a true antecedent and false consequent. The antecedent of a conditional specifies conditions that are supposed to be sufficient, in some sense, for those specified by the consequent; accordingly, no sentence can count as a true conditional of any sort if the former conditions obtain but the latter do not. The antecedent of (1), however, is simply (2), which is in fact true, and the consequent of (1) is (3), which we have agreed is false; hence, (1) is a false conditional.[7] Therefore, the inference from (1) and (2) to (3) is not a counterexample to modus ponens.

It follows that McGee's example does not threaten modus ponens even if we take the constituent conditionals as strong conditionals. Although I agree that in the relevant circumstances, we are apt to accept (1) and (2) and reject (3), I do not think that we are prepared to accept a conditional known to have a true antecedent and false consequent. According to an influential suggestion made by F. P. Ramsey,[8] one should evaluate a conditional by reflecting on certain conditional beliefs: add the antecedent of the conditional as an assumption to your current stock of beliefs, making whatever minimal revisions are necessary for consistency, and then evaluate the consequent in light of this adjusted set of beliefs. The leading idea, which has considerable plausibility, is that one should accept, or reject, a conditional just in case one is prepared to accept, or reject, the consequent under this hypothetical set of beliefs. Commentators have pointed out the shortcoming that it is difficult to specify a plausible set of constraints that will, in general, determine a unique way of adjusting one's initial beliefs to accommodate the assumption of the antecedent (especially when the antecedent is inconsistent with those beliefs). However this may be, it is not difficult to see that when we already accept the antecedent of a conditional, our adjusted stock of beliefs will be exactly the same as our initial stock of beliefs. Consequently, the question of whether one should accept, or reject, some conditional, when one already accepts the antecedent, is exactly the same as the question of whether one should accept, or reject, the consequent of that conditional. We see, accordingly, that (1) and (3) must have the same truth value, given the truth of (2), and that it would be irrational for someone who recognizes the truth of (2) to accept the former and not the latter. This shows, I think, that the conditional that we are prepared to accept, when we accept (1), does not have (3) as a false consequent and, thus, is not a strong conditional.

---

[7] Other considerations point to the same conclusion. The sentence (1′) 'If "A Republican will win the election" is true, then "If it is not Reagan who wins it will be Anderson" is true', which is logically equivalent to (1), says that, if (2) is true (which in fact it is), then (3) is true (which in fact it is not). How can the assumption that some sentence is true, when it actually is true, have any bearing on the truth value of a sentence that is false? If a sentence is in fact false, then it is false given any and every truth; so if (3) is false, as we have been assuming, then it is false given that (2) is true. But the claim that (3) is false if (2) is true is inconsistent with the claim that (3) is true if (2) is true, which is what (1′) says, hence, (1′) is false, in which case so is (1).

## IV

McGee's example, it turns out, points to a defect in the law of exportation rather than modus ponens. As we saw earlier, the law of exportation holds only for material conditionals; as we see now, the inference from (4) to (1) is a clear counterexample to the law. Accordingly, we have ample grounds for concluding that the law of exportation is not valid for strong conditionals. But even if we are agreed that the law of exportation does not hold for strong conditionals, we may be puzzled that it breaks down in this way. In fact, one may wonder why the argument presented earlier in support of (4) does not apply to (1) as well: in the relevant circumstances, there are exactly two Republican candidates, Reagan and Anderson. We begin with the assumption that a Republican will win the election; we then add the assumption that Reagan will not win. It seems reasonable to infer, in the presence of these assumptions, that Anderson will win the election. Accordingly, we concluded that (4) is true. Why does this reasoning not speak in favor of (1) as well as (4)?

---

[8] "General Propositions and Causality," in his *Foundation of Mathematics and Other Logical Essays* (New York: Humanities, 1950), p. 248. Stalnaker's account in "A Theory of Conditionals" is presented in part as a development of Ramsey's proposal.

Using considerations introduced earlier, we can explain why the reasoning does not apply to (1) and, hence, why exportation is invalid. If we examine (4) and (1) employing a model for evaluating conditionals like the one suggested by Ramsey, we shall see that while (4) and (1) introduce the same assumptions, they do not introduce them in the same way. To evaluate (4), we must adjust our initial stock of beliefs—which includes the beliefs that there are two Republican candidates, Reagan and Anderson, that Reagan will win, that a Republican will win, and that Anderson will come in a distant third—to make room for the assumption that a Republican other than Reagan will win. Accordingly, we must discard the belief that Reagan will win and make other modifications as well, but one belief we are not at liberty to drop is the belief that a Republican will win: obviously, we are unable to add the assumption that a Republican other than Reagan will win unless we retain our belief that a Republican will win. Thus, we must evaluate the consequent of (4) in light of the belief that a Republican will win, together with the other beliefs in our adjusted set, in particular, that the only other Republican candidate is Anderson. In order to evaluate (1), on the other hand, we must first look at the consequent of (1), that is, (3), in light of our initial stock of beliefs adjusted to include the antecedent of (1), that is, (2); of course, since we already accept (2), our adjusted stock of beliefs will be exactly the same as our initial stock of beliefs, which is why (1) and (3) have the same truth value.

But now someone may wonder why (3) turns out false, for it may appear that when we add the antecedent of (3) to our stock of beliefs, the resulting hypothetical set of beliefs will be the same as the one we had when we added the antecedent of (4) to our initial stock of beliefs. It may appear, in other words, that we evaluate the consequent of (3) and the consequent of (4) under the same hypothetical set of beliefs. But then, since the consequent of (3) and the consequent of (4) are one and the same, it would not be possible for (4) to be true and (3) false.[9]

---

[9] More generally, it *may appear* that when we evaluate the consequent of a conditional of the form ⌜If $\phi$ and $\psi$, then $\chi$⌝ and the consequent of the consequent of ⌜If $\phi$, then if $\psi$, then $\chi$⌝, we do so under the same hypothetical set of beliefs, namely, our initial stock of beliefs adjusted to include both $\phi$ and $\psi$, in which case the truth of the former would ensure that of the latter and the law of exportation would be sustained.

It is this line of thought that gives the law of exportation its illusion of cogency; it is, however, based on a mistake. To evaluate (3), we must adjust our stock of beliefs—which, again, includes the beliefs that there are two Republican candidates, Reagan and Anderson, that Reagan will win, that a Republican will win, and that Anderson will come in a distant third—to make room for the assumption that Reagan will not win. We can accommodate this assumption in one of two ways, by abandoning the belief that a Republican will win or by abandoning the belief that Anderson will come in a distant third. The former alternative was not available in the case of (4), for the antecedent of (4) includes the assumption that a Republican is going to win. But (3) introduces no such assumption; the antecedent of (3) says only that Reagan will not win. So it is open to us, in the case of (3), to incorporate the assumption that Reagan will not win by dropping the belief that a Republican will win. In fact, since the likely reason that we believe that a Republican is going to win (in the circumstances under consideration) is that we believe that Reagan is going to win, it is natural to think that the smaller revision in our set of beliefs would involve modifying the belief that a Republican will win, that is, (2), rather than modifying the belief that Anderson will be a distant third in the election, which is why we are apt to think that (3) is false. But then, so is (1).

In summary, the assumption that the law of exportation is valid for the conditionals occurring in the inference from (1) and (2) to (3) may lead us to suppose that the first premise is true and that the inference counts against modus ponens. The law of exportation, however, is valid only for truth-functional conditionals. Taken as a truth-functional conditional, the first premise is true, but then so is the conclusion; taken as a non-truth-functional conditional, the conclusion is false, but then so is the first premise. We see, accordingly, that the inference from (1) and (2) to (3) does not count against modus ponens on any consistent interpretation of the conditionals occurring in it.

BERNARD D. KATZ

University of Toronto

## COMMENTS AND CRITICISM

# SEMANTICS SAN DIEGO STYLE*

Ever since David Rummelhart and James McClelland[1] began their pioneering research into *parallel distributed processing* (PDP) over a decade ago, the University of California/San Diego has been a center for research into PDP models of cognition and the application of this research to traditional issues in such fields as anthropology, linguistics, philosophy, and psychology. As part of the interdisciplinary cognitive science program at UCSD, Paul M. Churchland[2] was quick to recognize and bruit the philosophic import of PDP, or connectionist, models of cognition, applying it initially to issues in the philosophy of mind and the philosophy of science. Most recently, Churchland has explored the possibility of using connectionist resources to defend a conceptual-role theory of semantics. In a recent series of exchanges, Churchland and the team of Jerry Fodor and Ernest Lepore[3] have been engaged in a debate over whether *state-space semantics* (SSS) can fortify conceptual-role semantics by providing a robust account of similarity of meaning. I agree with Churchland that SSS constitutes a "serious alternative"[4] to the *language of thought* (LOT); however, by construing SSS as a ver-

---

* I am grateful to Paul Churchland, Mark Collier, Adrian Cussins, Aarre Laakso, and Eric Thomson for helpful discussions on these issues and for comments on earlier drafts of this paper. I am also thankful for helpful comments from audiences at the University of California/San Diego and Rutgers University, where earlier drafts of this paper were read.

[1] *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Cambridge: MIT, 1986), which is considered the *locus classicus* of connectionism.

[2] *A Neurocomputational Perspective* (Cambridge: MIT, 1989).

[3] Fodor and Lepore, "Paul Churchland: State Space Semantics," in their *Holism: A Shopper's Guide* (Cambridge: Blackwell, 1992), pp. 187-207; Churchland, "Fodor and Lepore: State-space Semantics and Meaning Holism," *Philosophy and Phenomenological Research*, LIII (September 1993): 667-72; Fodor and Lepore, "Reply to Churchland," *Philosophy and Phenomenological Research*, LIII (September 1993): 679-82; Churchland, "Second Reply to Fodor and Lepore," in R. McCauley, ed., *The Churchlands and Their Critics* (Cambridge: Blackwell, 1996), pp. 278-83; Churchland, "Conceptual Similarity across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered," this JOURNAL, XCV, 1 (January 1998). 5-32.

[4] After showing that language of thought commits one to the view that we have not learned every language that we know, Fodor writes: "I admit that these conclusions really may seem scandalous. I should be inclined to view them as a reductio ad absurdum of the theory that learning a language is learning the semantic properties of its predicates, except that no serious alternative to that theory has ever been proposed"—*The Language of Thought* (Cambridge: Harvard, 1975), p. 82

0022-362X/99/9608/416–29

sion of conceptual-role semantics, I think that Churchland has actually done a disservice to the theory.

I make a distinction here between what I call the *vehicle question* and the *content question*. One of the central claims of this discussion is that SSS, properly construed, is a thesis about the vehicle of content, and thus is neutral with respect to any particular answer to the content question (for example, conceptual-role, causal, teleological).[5] Even if this is right, the question as to whether SSS can make conceptual-role semantics any more plausible is interesting and worth pursuing. The primary criticism that Fodor and Lepore levy against SSS as a version of conceptual-role semantics is that, like all other conceptual-role theories, it cannot provide a robust account of similarity of meaning. In his most recent paper, Churchland[6] argues that there exists a definitive response to this criticism. I argue here that the results to which Churchland appeals do not, in fact, provide an answer to Fodor and Lepore's challenge. I make a distinction between three different "stages" of SSS, and argue that Churchland takes a step in the right direction by making it clear that he is not defending the naive stage-one SSS attributed to him by Fodor and Lepore. Even the more sophisticated stage-two SSS which Churchland defends, however, does not come to grips with the full force of the similarity burden posed by Fodor and Lepore. I finally argue that a kind of conceptual-role SSS can meet the similarity burden if one moves to stage-three SSS, though in this version the extent to which conceptual role determines content is somewhat mitigated from that of a full-fledged conceptual-role theory of semantics.

## I. A THESIS ABOUT THE VEHICLE OF CONTENT

SSS is a thesis about the vehicle of semantic content. What I mean by a 'vehicle' of content is simply what carries the content. Another way to think of the vehicle of content for a system is that its properties are those properties which affect the computational functioning of that system.[7] I shall further explicate this conception of SSS here by contrasting it with two competing answers to the vehicle question: the *linguistic tradition* and the *mentalistic tradition*. While SSS is a competitor of LOT, which belongs to the latter tradition, I shall attempt to show that the two theories are actually closer to one another than either Fodor and Lepore or Churchland seem willing to acknowledge.

---

[5] Because of this, it has been suggested to me that I should call the position "state-space syntax." I am sticking with the label 'state-space semantics' both for historical reasons and because, even though the particular view of reference one accepts in the end may differ from one state-space theory to the next, they are all semantic theories within a state-space framework.

[6] "Conceptual Similarity."

[7] See Adrian Cussins, "The Connectionist Construction of Concepts," in Margaret Boden, ed., *The Philosophy of Artificial Intelligence* (New York: Oxford, 1990), 368-440, here p. 369

The view that stands in sharpest contrast to SSS is what I am calling the linguistic tradition, a view that holds the following two claims:

(1) Linguistic entities carry content independently of how they connect with any particular mental representation.
(2) The computational properties of our "syntactic/semantic engine" are those exhibited by language (for example, the inferential rules of propositional logic).

For those who hold (1), linguistic entities are the primary unit of meaning, and the task in learning a language is to get one's mental representations to correspond to the right words. On this picture, words must hook onto mental representations which, in turn, must hook onto the world. I take it that it is because the primary reference relation on this view is from words to mental representations (rather than to the world) that Hilary Putnam in his Dewey Lectures[8] argues that the problem of reference is simply a replay of the seventeenth-century empiricist problem of perception. The second of these claims follows naturally from the first. If the mind has content in virtue of corresponding to certain linguistic entities, then the rules that govern the computational functioning of linguistic entities, namely, propositional logic, must also be the rules that govern the computational functioning of the mind.

An alternative tradition, which can be called the mentalistic tradition, turns the picture of claim (1) on its head. That is, rather than conceive of language learning as a matter of getting one's mental representations to correspond to words, one could see language learning as a matter of getting words to correspond to one's mental representations.[9] The most systematically developed contemporary account of this alternative is probably LOT, which holds that the vehicle of content is a mental representation, in particular symbols in a "language of thought."[10] These symbols are the semantic simples of the language of thought, like propositions are the semantic simples of propositional logic. While LOT is a member of the mentalistic tradition, in holding (2) I take its defenders still to have a foot in the linguistic tradition, for they still view thought as proceeding in the same way language does; that is, they take the computational properties of mental representations to be the same as the computational properties of propositions.

---

[8] "Sense, Nonsense, and the Senses: An Inquiry into the Powers of the Human Mind," this JOURNAL, XCI, 9 (September 1994): 445-517.

[9] On this view, the primary reference relation is from mental representations to the world. Since the primary reference relation already involves the world, this may be a way around the problem that Putnam presents in his Dewey Lectures (*ibid*).

[10] See Fodor, *The Language of Thought*.

SSS follows LOT in rejecting the first claim of the linguistic tradition. In this respect, SSS and LOT can both be seen as members of the same tradition in the philosophy of language. Both SSS and LOT start with a representational theory of mind, that there must be "mental particulars possessed of semantic *and syntactic* properties"[11] and that "mental processes are transformations of mental representations."[12] In addition to a commitment to realism about mental representation, SSS shares with LOT the conviction that learning a natural language requires having a system of mental representations in place. That is, SSS agrees that "certain kinds of very central patterns of psychological explanation presuppose the availability, to the behaving organism, of some sort of representational system,"[13] and especially that "to extend our psychology to infrahuman species is thereby to commit ourselves to cognitive processes mediated by representational systems other than natural languages."[14] Because LOT assumes that mental representations are symbols in a *language* of thought, it insists that the computational properties of mental representations follow that of natural language. This is where SSS parts company with LOT. Because SSS has a different account of mental representation, it has a different account of the computational properties of those representations.

SSS is based on a connectionist theory of mind. The primary unit of representation in connectionist models of cognition is a *pattern of*

---

[11] Fodor, "Fodor's Guide to Mental Representation," in his *A Theory of Content and Other Essays* (Cambridge: MIT, 1990), pp 3-27, here p. 23.

[12] In his "Guide to Mental Representation," Fodor denies that Churchland holds a representational theory of mind because Churchland is an eliminativist about the mental states postulated by folk psychology. It would only follow from Churchland's eliminativism that he rejects the representational theory of mind if the mental states of folk psychology were the only kind of mental representations possible. But this is precisely what any connectionist theory of mind denies. Connectionism maintains that mental representations are best thought of as multidimensional state-spaces. (This does not mean that connectionism is incompatible with folk psychology, only that one can eliminate the entities of folk psychology and still be left with mental representations.)

[13] *The Language of Thought*, p 31

[14] Churchland takes a similar line for the case of children when discussing what he labels "the problem of early stages"—*Scientific Realism and the Plasticity of Mind* (New York: Cambridge, 1979), pp 127-37.

*activation* across a population of neurons, such as V1 of the primary visual cortex, or a certain layer in an artificial network. This pattern of activation can be thought of as an *n*-dimensional *activation vector*, where *n* is the number of neurons in the population, as it can be uniquely characterized by an *n*-tuple of numbers with each element corresponding to the activation level of a given neuron. The pattern of activation can also be represented as a *point* in an *n*-dimensional state-space, with each axis of the space corresponding to each neuron in the population. One can also combine the last two and represent the pattern as a *vector in state-space*. All of these are merely different ways of representing the same thing. According to SSS, the vehicle of content = a pattern of activation = an activation vector = a point in state-space = a vector in state-space.[18] Vectors in state-spaces have computational properties different from those of symbols in a language of thought. In fact, the kinematics and dynamics of thought processing on the state-space model looks nothing like that of language. According to SSS, thought proceeds by vector-to-vector transformations rather than by logical inferences between sets of sentences.

I have claimed that SSS properly construed is a thesis about the vehicle of content. In claiming that the vehicle of content is the state-space, I take SSS to be rejecting both claims of the linguistic tradition and instead holding the following two claims:

(1′) Mental representations (in particular state-spaces) are the vehicles of content, and linguistic entities have content only insofar as they "hook onto" these mental representations.

(2′) The computational properties of our "syntactic/semantic engine" are those of connectionist models of cognition (for example, vector-to-vector transformation and prototype activation).

In claiming that SSS is a thesis about the vehicle of content, I am also claiming that SSS is neutral with respect to the "content question," the question as to how a particular vehicle obtains its content. One could simply apply her favorite semantic theory to SSS. For example, a causal role SSS would state that a given activation vector *v* in *S* means (refers to) *X* if and only if: (a) '*X* causes *v*' is a law, (b) at least some *X*s cause *v* (in the actual world), and (c) for all $Y \neq X$ which cause *v*, *Y* would not have caused *v* were either (a) or (b) false; a functional/teleological SSS would state that a given activation vector *v* in *S* means (has the content of) *X* if and only if it is the function of *v* to indicate *X*s to the system in which *S* is instantiated; a conceptual-role SSS would state that the content of a given activation vector, *v*, in a state-space, *S*, is determined by the role it plays in the larger conceptual economy of the agent in which *S* is instantiated; and so on for any semantic theory. In making this last move, I have conceded to Fodor and Lepore that in order to ascribe content to a particular vector, one must already presuppose some *semantic* theory. But this is not an argument against SSS as a thesis about the vehicle of content. In fairness to Fodor and Lepore, they were not arguing against SSS so construed and thus did not intend the objection to be directed at this conception of SSS. It may be the case that I have said nothing with which a defender of Fodorian style *semantics* would disagree (see footnote 5). In claiming that one can apply *any* semantic theory to SSS, I have implicitly asserted that Churchland's conceptual-role version of SSS can work; however, that is precisely what the primary debate has been between Churchland and Fodor and Lepore, and it is to this debate that I now turn in the remainder of this discussion.

---

[18] While in his "Conceptual Similarity" Churchland still sees SSS as a version of holism, he does emphasis what the vehicle of content is for SSS, and provides even more detail, including helpful diagrams, than that given here.

## II CONCEPTUAL-ROLE STATE-SPACE SEMANTICS

The primary criticism that Fodor and Lepore levy against SSS as a version of conceptual-role semantics is that it falls victim to the same criticism that afflicts all conceptual-role theories, namely, that it cannot provide a robust account of similarity of meaning. In his most recent paper, Churchland argues that recent results in cognitive science "constitute a decisive answer to Fodor and Lepore's challenge."[14] I shall argue here that the results to which Churchland appeals do not, in fact, get conceptual-role SSS off Fodor and Lepore's hook.

Before moving to Churchland's defense, it will be instructive to revisit the original challenge facing conceptual-role SSS. The problem was that similarity of meaning between two state-spaces is understood as similarity of activation across dimensions. This requires a way of distinguishing when two spaces are two tokens of the same state-space type, which in turn requires that the content of each dimension in one space be identical to the content of its corresponding dimension in the other. In order for this condition to be met, there must be a property expressed by each dimension in the respective spaces. But if one is to attribute a property to the dimensions, then one must already have a notion of meaning. Furthermore, in order to tell whether the dimensions of each space express the *same* property, one must already have a notion of identity of meaning. Not only does SSS not solve the problem of similarity of meaning, it fails even to avoid the problem, and so cannot make conceptual-role semantics any more plausible than it was to begin with.[17]

---

[14] "Conceptual Similarity," p. 5.

This criticism was originally aimed at what I shall call *stage-one* SSS, which individuates the content of activation vectors by absolute position in state-space (that is, position in relation to the axes). It thus attempts to provide a robust account of similarity of meaning by taking similarity of meaning as equivalent to similarity of location in state-space. There are passages from Churchland[18] which suggest that this is the conception of SSS which he holds:

> ...a visual sensation of any specific color is literally identical with a specific triplet of spiking frequencies in some triune brain system...then the similarity of two color sensations emerges as just the proximity of their respective state-space *positions* (*ibid.*, p. 104; my italics).

As the above passage shows, this account was originally intended as an account of experience, but, as Fodor and Lepore recognize, the state-space model can provide an account of semantic content as well as of the content of phenomenal experience. That is, not only can one consider the *experience* of red to be a pattern of activation (that is, a point in state-space), but one may also consider the *semantic* content of the term 'red' to be a pattern of activation.

The above criticism is sound against stage-one SSS. Neither I nor Churchland, however, wish to defend stage-one SSS. In his more recent papers on the topic, Churchland clarifies that he is not defending stage-one SSS, but rather (call it) *stage-two* SSS. While stage-one SSS ascribes content to vectors in virtue of the content of the dimensions and defines similarity across spaces in terms of similar absolute position in state-space, stage-two SSS holds neither of these views. In the previous section, I showed how conceptual-role SSS (as well as any other version of SSS) can ascribe content to activation vectors without ascribing content to the individual dimensions of the space. With regard to similarity, stage-two SSS defines similarity across spaces in terms of relative position, that is, the position of a vector in relation to the other vectors. Stage-two SSS then individuates the content of activation vectors by similar relative positions, that is, by similarity of clustering. The basic idea behind the similarity of clustering condition is that a given activation vector $v$ in space $S_1$ is similar in content to an activation vector $w$ in space $S_2$ just in case $v$ stands in the same set of relations to the other vectors in $S_1$ as does $w$ to the other vectors in $S_2$. For stage-two SSS, vectors of different spaces can be representationally similar to one another regardless both of the content of the dimensions of each space and of the number of dimensions of each space.[19]

---

[17] Fodor and Lepore, "State-Space Semantics," pp. 197-99.

[18] "Some Reductive Strategies in Cognitive Neurobiology," in his *A Neurocomputational Perspective*, pp. 77-110.

With stage-two SSS formulated, we can now address the question of whether Fodor and Lepore's criticism against conceptual-role SSS still has force. The burden for the proponent of conceptual-role SSS is to provide an account of how to determine whether the activation vectors of one state-space stand in the same set of relations to one another as do the activation vectors of another space. If the proponent of conceptual-role SSS can meet this burden, then indeed SSS has made the world safe for conceptual-role semantics by providing a robust account of similarity of meaning.

Churchland argues that the experimental results of the work by Aarre Laakso and Garrison Cottrell[20] "constitute a decisive answer to Fodor and Lepore's challenge." Laakso and Cottrell have developed a measure similar to Guttman's point alienation, which purports to measure the degree of similarity between two representational structures (for example, two state-spaces). I shall not here go into detail about the experiment, as a fairly detailed account was presented by Churchland. The basic idea behind the measure is that one can look at three points at a time, $r$, $p$, and $q$, hold one of the points, say, $r$, as the anchor, and compute whether $p$ or $q$ is closer to $r$. One then does the same thing for the corresponding points $r'$, $p'$, $q'$ of another space and compares the ranking. The procedure is then repeated for all possible triplets. If the two spaces have exactly the same set of rankings, the measure will have a value of 1, and if they have exactly the opposite set of rankings, the measure will have a value of -1.

Laakso and Cottrell used this measure to compare networks that were trained up to identify an input pattern as being one of five colors: red, yellow, green, blue, or purple. They trained up twenty different networks, five networks for each of four different ways of representing the input pattern.[21] They then used the measure to compute the rank-order correlation both between networks that were trained up on the same input-encoding and between networks that were trained up on different input-encodings. They found that the rank-order correlation between networks trained up on the same input-encoding was almost always greater than .95, meaning that the activation vectors of the different networks were very similarly clustered between states. They also found that the rank-order correlation between networks trained up on different input-encodings was also very high, usually greater than .9 (with a minimum value of .85). These results show that even when networks have completely different ways of representing an input pattern at the "sensory" level, they can still have highly similar ways of representing that input at a higher level (that is, of "conceptualizing" the input). For this experiment, the middle-layer activation vectors of the state-space of one network do stand in approximately the same set of relations to each other as do the activation vectors in the state-space of another network.

---

[19] In his "Conceptual Similarity," pp 8-13, Churchland develops this idea in more detail and provides some very helpful diagrams, including one of a "family-space" consisting of the Hatfields, McCoys, and Andersons.

[20] "Qualia and Cluster Analysis. Assessing Representational Similarity between Neural Systems" (under review), presented in earlier draft to the 1996 meeting of the Society for Philosophy and Psychology, San Francisco.

[21] The four ways of representing the inputs were: (1) an 8-bit binary number, (2) a decimal number between 0 and 1, (3) a gaussian curve over five elements, and (4) sequentially as a 3-digit decimal. So each network had a different number of cells (that is, dimensions) and the cells of one network represent completely different properties than those of another.

It is easy to see why Churchland took these results to be a conclusive answer to Fodor and Lepore, as it seems clear that it shows that there exists an empirical method for determining whether the activation vectors of two spaces are similarly clustered, and thus that the burden for conceptual-role SSS as laid out above is met. The question, however, is whether it does so in a way that does not already presuppose a theory of content, for the original charge against conceptual-role SSS was that it could only provide a robust account of similarity of meaning by presupposing a theory of meaning.

The SSS skeptic will probably have already noticed that the Laakso-Cottrell measure involved comparing a triad of vectors in one space with a "corresponding" triad of vectors in a different space. That is, the measure presupposed that there was no problem with labeling one vector $r$ in space $S_1$ and another "corresponding" vector $r'$ in $S_2$. In the actual experiment, this in fact was not a problem, for which vectors corresponded to which was simply determined by the input pattern that gave rise to the vectors. If two networks were both shown input pattern $\text{red}_3$, then the activation vector produced at the hidden layer in one network is labeled $r_3$, and the activation vector produced at the hidden layer of the other network is labeled $r_3'$, and similarly for, say, $\text{orange}_2$ and $\text{blue}_1$. The two triads can then be put through the measure to see if the two networks conceptualized $\text{red}_3$, $\text{orange}_2$, and $\text{blue}_1$ similarly (and likewise for all such triads of inputs).

This may be fine as an account of the similarity of qualia clustering; however, if the issue is how one goes about assigning content to vectors in the first place, then the Laakso-Cottrell measure does not seem to be of help to conceptual-role SSS. Consider again the initial charge against SSS:

> We're claiming, in effect, that Churchland has confused himself by taking the *labels* on the semantic dimensions for granted. The label on a dimension says *how positions along the dimension are to be interpreted*; for example, it says that they're to be interpreted as expressing degrees of F-ness. To label a dimension as the F-ness dimension is thus to invite the question "in virtue of what do the values of this dimension express degrees of F-ness rather than, say, degrees of G-ness?"[22]

If the proposal is to give an account of content similarity by appealing to position in state-space, then the proponent of SSS has given a circular account by helping himself to semantic labels for the dimensions, which means he has already presupposed a theory of meaning.

The initial response from Churchland was to deny that it is necessary to label the dimensions with semantic content in order to ascribe content to a vector:

> What makes each vector, Patricia's and mine, a "kitten" vector is not the identity of our respective patterns of neuronal activity across the hidden layer (these are likely quite different), nor the semantic identity of the constituting dimensions in her hidden-unit population and in mine (their diffuse "contents" may well be quite idiosyncratic). What gives this vector the content "kitten" is the overall role that this vector plays in the larger cognitive and motor economy of which it is an interlocking part.[23]

This, though, only partially answers Fodor's attack. It does show that it is not necessary to ascribe content to the individual dimensions of a space in order to ascribe content to a particular vector in that space (this is the move from stage-one to stage-two SSS). But it does help itself to a pre-existing theory of meaning, namely, conceptual-role semantics. I have already noted that this is not an argument against SSS conceived as a thesis about the vehicle of content; however, the issue here is whether SSS makes conceptual-role semantics any more plausible (than, say, a sententially based conceptual-role approach), and at stage-two SSS, the issue of content similarity simply arises anew. The question is how one can determine whether Paul's 'kitten' vector plays a similar role "in the larger cognitive and motor economy of which it is an interlocking part" as does Patricia's 'kitten' vector.

---

[22] Fodor and Lepore, "State-Space Semantics," p. 199.

[23] Churchland, "State-Space Semantics and Meaning Holism," p. 670

I argued above that under stage-two SSS this challenge can be met if one can provide a criterion for determining whether the vectors in one space stand in the same set of relations to one another as do those of another vector. The next response by Churchland was that the Laakso-Cottrell results meet this challenge. If this only works by first labeling the vectors of each space, however, then Churchland has cheated even worse than was initially charged; for he has helped himself not simply to labels for the dimensions, but to labels for the actual vectors themselves. If there was no problem about when one could say that a vector of one space corresponds to a vector of another space, then why is the Laakso-Cottrell measure needed in the first place? To solve the problem in the way that was done for the actual experiment and say that the corresponding vectors were identified by corresponding input is just to help oneself to an externalist theory of content. But to appeal to some externalist theory to fix the content of vectors would obviate the conceptual-role part of the content story.

This does not mean that conceptual-role SSS is a doomed theory. It just shows that the Laakso-Cottrell results does not get conceptual-role SSS out of the similarity of meaning problem. There may be as-of-yet-untapped resources that meet the similarity of clustering burden without presupposing some externalist theory of content. This also does not show that conceptual-role semantics is bankrupt. All this means is that SSS has not yet made conceptual-role semantics any *more* plausible than it was to begin with. This will be cause for little concern for those who think that conceptual-role semantics is already the most plausible semantic theory in the offing. I also do not mean to suggest that the Laakso-Cottrell results are of no use to SSS. In the next section, I shall develop stage-three SSS and show the importance of these results for that conception.

## III. STAGE-THREE STATE-SPACE SEMANTICS

While the foes of conceptual-role semantics are now smiling, the friends of conceptual-role semantics may be skeptical of my abrupt dismissal of the Laakso-Cottrell results. After all, I suggested in the last section that, if one already had labels for the vectors of each space, then the results of applying the Laakso-Cottrell measure were, at best, redundant and, at worst, completely irrelevant. But, one may respond on behalf of conceptual-role SSS, the results could have been negative! That is, once the Laakso-Cottrell measure was applied to the conceptual structure at the middle layer of the differing networks, a low or negative score could have been the result. In fact, when the measure was applied to the conceptual structure at the input layer, the results were negative—the networks did not conceptualize the colors similarly at all at the input level. So the positive results at the middle layer must mean something. I shall propose here that the way to resolve the tension between the conclusion of the last section and the intuition that the positive results of the Laakso-Cottrell experiment mean something is to draw a distinction between two types of representational content, which I shall call *correlational* and *embedded* content.

The correlational content of a representation is roughly that with which it is correlated. If a given representation is correlated with cows, then the correlational content of that representation is 'cow'. This notion is very close to what the Churchlands called "calibrational content."[24] What I mean by correlational content is also very close to the ordinary notion of reference. The main difference is that (depending on one's theory of reference) the two notions may not be co-extensional. It is probably the case that the correlational content of a given representation is always that to which it refers, though there could be cases in which a representation is nonreferring but still has correlational content. One such example might be fictional characters. The name 'Pegasus' may be nonreferring, but it could still have correlational content, namely, a certain creature from Greek mythology. Divergence between reference and correlational content may also arise when one's ontology does not correspond to the world; the term 'demonic possession', for example, may be nonreferring and yet have correlational content, perhaps a particular type of neurosis.

The content that I want to designate by the label 'embedded content' is that which a representation has in virtue of being embedded in a certain cognitive structure of a system (or psychological reality of an agent).[25] That is, the embedded content of a given representation is the significance that its correlational content has for the cognitive economy of the system of which it is a part, given the cognitive structure of that system. This will include both the relations that the representation bears to other representations in the system and to its downstream causal connections such as its effect on thought (mental processing) and behavior (motor control). My notion of embedded content is very similar to the Churchlands' notion of translational content, which "is a matter of the inferential/computational relations it bears to all the rest of [the system's] representations."[26] One main difference between my embedded and their "translational" content is that they do not seem to include the behavioral effects of a representation as part of the translational content (though perhaps they would not have a problem with doing so). Even if there is no conceptual difference between the two notions, one problem I have with the label 'translational' is that it sounds too conceptual and suggests a close connection to natural language, whereas, the content that I want to designate with the label 'embedded' need not be conceptual, nor need it only apply to the content of terms in a natural language.

---

[24] Paul Churchland and Patricia Churchland, "Functionalism, Intentionality, and Qualia," in Paul Churchland, *A Neurocomputational Perspective*, p. 43.

[25] I specifically want to avoid calling embedded content *conceptual*, because I want to allow for embedded content to be nonconceptual, in the sense specified by Gareth Evans, *The Varieties of Reference* (New York: Oxford, 1982); or Cummins.

Given this distinction, there is a straightforward way of resolving the tension outlined above. The correlational content of a given activation vector is best specified on some externalist approach. This is how one can label the vectors of a given space. I have tried in specifying what I mean by correlational content not to prejudice the issue in favor of externalism, but instead to leave room for an internalist (for example, conceptual-role) approach to correlational content. Such an approach, however, will have to deal with the similarity argument with which we began, and one will not be able to appeal to the Laakso-Cottrell results to meet this challenge with respect to correlational content. An internalist approach to correlational content will also have to deal with the many arguments against internalist approaches to reference. Embedded content, on the other hand, does fit most naturally with some internalist, conceptual-role approach. The Laakso-Cottrell measure then provides a way of determining whether (at least one aspect of) the embedded content of a vector with a particular correlational content in one space is similar to that of a vector with the same correlational content in another space. The positive results of the Laakso-Cottrell experiment show that the embedded content of the activation vectors whose correlational content is a certain color was indeed similar across the state-spaces at the middle-layer level of differing networks. In solving the similarity of content problem for embedded content, one can make use of the Laakso-Cottrell measure without circularity.[27]

---

[26] "Functionalism, Intentionality, and Qualia," p. 43.

[27] The distinction between correlational and embedded content may also provide a framework for addressing the problems surrounding the connection between certain kinds of beliefs (for example, indexical or moral) and action. For example, while it is true that the correlational content of indexical representations is necessarily specified token reflexively, the substantive question is whether the embedded content of indexical representations necessarily contains a motivational force (see my "What Is Essential about Indexicals?" forthcoming in *Philosophical*

## IV CONCLUSIONS

All versions of SSS have in common the claim that the occurrent vehicle of content is an activation vector across a population of neurons. Stage-two SSS differs from stage-one SSS in how it individuates the content of those vectors, but both versions assume a unitary notion of content. That is, both versions implicitly assume that activation vectors are the vehicles for a single type of content. *Stage-three* SSS differs from the first two versions in holding that the content of vectors must be individuated along (at least) two dimensions. Churchland correctly identifies the vehicle of content for SSS, and as such is an improvement over LOT. He does a disservice to the theory, however, by presenting it specifically as a version of conceptual-role semantics. In meeting Fodor and Lepore's similarity of meaning challenge for conceptual-role SSS, Churchland again makes a move in the right direction by distinguishing stage-two from stage-one SSS; however, the experimental results to which Churchland appeals do not, in fact, get conceptual-role SSS off Fodor and Lepore's hook. By moving to stage-three SSS, one can allow conceptual-role to determine one type of content and use the Laakso-Cottrell results to provide a robust account of the similarity of that type of content. In the end, though, regardless of the extent to which conceptual-role provides an answer to the content question, SSS, properly construed as a thesis about the vehicle of content, is plausible and Churchland has done a service by bringing its existence to light.

EVAN TIFFANY

University of California/San Diego

---

*Studies*). Likewise, the debate about moral motivation can be construed as whether the embedded content of moral representations necessarily contains a motivational force. It may be the case that much of the dispute between internalists and externalists boils down to a conflation of the two types of content. That is, perhaps internalists of an expressivist or prescriptivist ilk may have been trying to provide an account of the embedded content of moral terms, while externalist naturalists may have been trying to provide an account of correlational content. The former, in asserting that the embedded content contains a motivational force, and the latter, in denying that the correlational content does, are not in fact disagreeing

## BOOK REVIEWS

*Having Thought: Essays in the Metaphysics of Mind.* JOHN HAUGELAND. Cambridge: Harvard University Press, 1998. 390 p. Cloth $29.95.

On the back of the dust jacket of this fine book, one can barely make out two representations of a customized penny for our thoughts, drawn by John Haugeland. Accompanying Honest Abe on the heads side appear the words 'Existential Commitment', 'Thought', and 'Self'; while tails shows the Lincoln Memorial and *E pluribus unum*, surrounded by two unlikely additions: 'Constituted Domain' and 'Objects'. Haugeland explains: "The basic Kantian/Heideggerian conclusion can be summed up this way: the constituted objective world and the free constituting subject are intelligible only as two sides of one coin." Not everything with eyes and ears and a brain is a "free constituting subject"; apes and dogs and dolphins do not have ontologies because they do not have thoughts; they do not have thoughts because they do not have the "censoriousness" in their cultures or social structures which provides the leverage for understanding, distinguishing a true thought from a false thought, and without that, thoughts cannot really have content. There are, to be sure, important differences in the *Umwelt* or manifest image of different species, but cats have no more metaphysics than clams or chrysanthemums do. Objects are constituted by people only, and our ultimately moral sense of norms, of "existential" commitment, far from being a sort of ethical add-on to the factual world of objects and properties, is the very ground on which our capacity to know, and reflect upon, objects depends.

This is a collection of essays, all but two previously published, so the question must be posed: Are the benefits of the synergy worth the cost? Here is one point in favor: I had read almost all these essays several times, and discussed some of them with Haugeland for dozens of hours, and even published commentaries on several of them, but I did not get the big picture until I read the book. That is, I had seen what he was driving at, but was unpersuaded until I saw the pieces put together anew. I now think he is largely right. Even better: the points on which I harbor residual doubts are, in general, more empirical than philosophical, and will mostly sort themselves out as we learn more about how minds are actually constructed.

0022-362X/99/9608/430–35

Haugeland claims that the difference in kind, not degree, between the "cognitive" capacities of people and animals is too obvious to require argument. This might be true in the philosophical community, but it is at any rate no longer true in the world of animal cognition research, where the backlash against the behaviorism of yore makes fighting words of any such pronouncement. Still, Haugeland is right that the gulf between animal mind and human mind is huge. It is not that surprising to see a philosopher claim that even the brightest of beasts do not have a manifest image much like ours, but the paths he takes from this starting point are strikingly novel. In particular, he firmly bucks the fashion and declares that language is *not* the key to this huge chasm.

> I certainly do not deny that language is characteristic of people, and centrally important to what we are. But I want to resist what strikes me as an all too prevalent tendency to make of language a magic fulcrum on which every human quality turns—as if language were the whole essence of what distinguishes us from machines on the one hand, and beasts on the other (256).

A strange centerpiece of Haugeland's case against the role of language in making—not just marking—the chasm is an extended thought experiment about imaginary "supermonkeys" who have no language but nevertheless play chess! They play real chess, embedded in the matrix of norms that constitute proper play, not the pseudo-chess of Deep Blue, whose *Umwelt* is as bereft of the appreciations of norms as a spreadsheet's *Umwelt* is of an appreciation of the human value of a dollar. There is nothing constitutively linguistic about chess, certainly, but the contortions of Haugeland's tale of how the supermonkeys come to their norm-guided involvement in the game are unpersuasive to me. Let us compare these curious beings to others we might imagine: superorangutans, also languageless, who somehow manage to build and navigate aircraft carriers. An aircraft carrier with no written signs, no instruction manuals, no gossiping crew or order-issuing captain may be *logically* possible in some philosophical sense, but so what? Haugeland is right to resist the popular, lazy idea of language as a magic fulcrum; but what about language as a garden-variety fulcrum, *causally* required in this world for chess, for building aircraft carriers—and for "normativity" itself? If human language and human norms emerged in a co-evolving package, with innovations in each laying the ground for elaborations in the other, back and forth—and that is how evolution typically works—then both 'Language made norms possible' and 'Norms made language possible' are misleading oversimplifications. Haugeland's resistance to the former is, however, salutary.

Haugeland's work on the conceptual foundations of cognitive science has been deservedly influential, and several of the classics are reprinted here: his auspicious debut with a target article in *Behavioral and Brain Sciences*, "The Nature and Plausibility of Cognitivism" (1978), and "The Intentionality All-Stars" (1990). Put these together with "Analog and Analog" (1981) and "Mind Embodied and Embedded" (1995), and you have a lucid and balanced setting for what I consider the most valuable essay in the collection, "Representational Genera" (1989). He describes this as "perhaps the most disheveled piece in the volume" (5), but while one can see why he was tempted to issue this self-disparagement, the fact is that this is a stunning piece, hugely ambitious, and surprisingly successful in meeting its ambitions. It purports to be a revisionist taxonomy of representations, and that it surely is, but its new taxonomy amounts to a strikingly new way of thinking about representations. Few topics have been mulled over more than representations in the last half century, and one would think that by now all the major moves had been made, and that it was time for fine tuning. Indeed, as Haugeland says, all this mulling has congealed into *canonical accounts*—which he nicely defines as "what almost everybody expects almost everybody else to believe."

According to the canon, the two main genera are logical/linguistic and iconic/pictorial. Today, we are faced with the interesting question of whether the much-touted *distributed representations* of connectionism (or holograms, for instance) deserve to be recognized as a distinct genus, as different from pictures and sentences as pictures and sentences are from each other. Although the quirky problems in characterizing the "essences" of these primary genera have long been acknowledged, the canonical wisdom has been that this is *not* to be done in terms of the content of what they represent, but somehow in their manner or format or system of representing. Haugeland sets out to show that the received wisdom is not just bedeviled by unsolved problems (everybody knows that), but that these problems are unlikely to be solved since this underlying assumption is a huge mistake. This certainly must strike the mavens of the field, as it first did me, as wildly implausible. Ah, all the more reason to think that maybe Haugeland has caught a big fish.

He develops his case with a very well-considered game plan. Indeed, this essay is a bravura display of philosophical move making of the highest order. If I were to put together a "how to" book on "how to do philosophy," this essay would be one I would dissect at length, revealing its virtues. Most philosophical books and articles have a rather plodding plan of attack, along the lines of: many believe that *p*; there are *n* familiar arguments against *p*, and there are *n* +δ familiar rebuttals to those arguments; here is a new argument against *p*. Haugeland sees his task more subtly in a three-stage development. It begins with some loosening-up counterexamples which he does not bother refining in the standard way, since they are just attention directors; this he follows by some "outlandish counter theses," which are themselves not to be taken seriously but which hint at the opening to come in the third movement: a way of seeing how our canonical assumption *might* be undercut.

> I am arguing instead that all three canonical accounts misfire for a single underlying reason–namely, they all mislocated the generic essences in the representing relation–and hence, the appropriate response is neither repair nor despair, but fundamental reconception (177).

The fundamental reconception involves taking on a hitherto unremarked distinction between *recording*, which can be a "witless" process, and *representing*, which, in general, cannot. This is promising, somewhat to our surprise, precisely because Haugeland is *not* planning to use it to make some 'No machine can ever do *x*' claim—a good example of the tactical discretion which abounds in this essay. His distinction between "witless" and "knowing" processes enables him to describe a new type of theoretical entity: "'virtual logical representations' (the 'facts')" (183). We all know, do we not, that any doctrine of atomic and molecular facts is doomed if it treats such facts as somehow independent of language expressing them? But once again, he is not planning to push his doctrine of facts into that hopeless cause; he has imagined another, more defensible use for them.

The task facing any "Heideggerian/Kantian" theorist is to do justice to the role of *us* in *constituting* the denizens of "our" world without lapsing into awful relativism/subjectivism on one side or caving in to noumena or a "God's eye view," on the other. Haugeland's solution, which grows on me, is to show how and why it is *hard* to "constitute" a world (that takes care of anything-goes relativism) but not because there is a privileged way that the world—the *real* world—has always been constituted. His view is a close kin, I think, of my view of the evolution of colors: before color vision evolved on this planet, sunsets and cliffs and volcanic eruptions had the reflective properties they did, but *it makes no sense* to ask if those sunsets were, say, red–since that question has no meaning independently of a reference class of normal observers. We can, of course, extrapolate back from our current vantage point and fix and answer such questions, using ourselves as the touchstone for colors, but we must recognize that we are doing that. Were there dinosaurs before *Homo sapiens* came along and invented censoriousness and then ontology so that dinosaurs could be constituted? Of course there were, but do not make the mistake of thinking that this acknowledges a fact that is independent of *Homo sapiens*.

Haugeland is not a neurophilosopher or psychophilosopher or cognitive scientist; he is a pure (or hard core—*chacun à son goût*) philosopher. Yet he brings to his task—the unflinching consideration of grand metaphysical themes from Immanuel Kant to Richard Rorty—an engineer's deep commitment to clarity and making sure things work. He once turned an old IBM selectric typewriter into a computer printer (think about *that* interface problem!) and is the author (in assembly language) of some impressive software for doing automatic file back-ups, so he has plenty of hands-on acquaintance with the relations between mechanisms and codes, "symbols" and realities, hardware and software. Like the computer scientist/metaphysician Brian Cantwell Smith, whose remarkable book *On the Origin of Objects*[1] would make a superb seminar mate with this volume, he has seen that philosophers' fantasies about computation and representation have not been innocent oversimplifications but major misleaders that must be replaced.

The corrections of outlook Haugeland recommends are always presented in a constructive spirit. "The Intentionality All-Stars" is a frequently hilarious view of all the occupied positions on the topic of intentionality, laid out on a baseball diamond, and while he makes it very clear where his own sympathies lie, there are no idiots on the field of play. Haugeland's pedagogical caricatures, unlike many in philosophy, bring out the best in each position. It turns out, for instance, that the main differences along the spectrum from first base (neo-Cartesian theory—for example, Jerry Fodor) and second base (neo-behaviorist—for example, Donald Davidson and me), to third base (neo-pragmatic theory—for example, John Dewey, Martin Heidegger, Wilfrid Sellars, and Haugeland himself) can be seen to be due mainly to different empirical hunches about the relative importance of factors admitted by all. For all its brilliance, this essay will not be readily understood by those without a substantial knowledge of baseball; the many allusions to the lore of the game really do a lot of work, and I know from experience that British philosophers of mind miss a lot of the wisdom. My prescription: philosophers everywhere now have a very good reason to add baseball to their stock of common knowledge. I also know from experience that my American undergraduates, who know the baseball, need help appreciating the essay's glories because they do not have the requisite familiarity with the philosophical literature it positions. Same prescription: do your homework; it is worth it.

---

[1] Cambridge: MIT, 1996.

Finally, I want to end this review with a specific compliment and a general complaint and call to arms. Some of the footnotes in this book are gems, deeply important, and not to be missed: for example, footnote 7 on page 44; footnote 10 on page 61; footnote 9 on pages 164-65. Down with "endnotes"! There is no excuse, in this age of computer typesetting, for the "economy" of endnotes, either at the end of each chapter or at the end of the book. Few if any philosophy books need be read with only one hand, but why should we need three? I find that placing a Post-it on the edge of the current notes page provides a useful fingertab, and can be moved along as one goes, but this is still an annoying interruption. Authors, join the cause: ask publishers to put *substantive* footnotes at the bottom of each page, and all bibliographical material in some version of scientific format (Jones, 1994, p. 17)! Then philosophy books can aspire to be page turners in the good sense.

DANIEL C. DENNETT

Tufts University

## NEW BOOKS

STROLL, AVRUM: *Sketches of Landscapes: Philosophy by Example.* Cambridge: MIT, 1997. xiv, 282 p. Cloth $35.00.

STURROCK, JOHN: *The World from Paris: Essays on Modern French Thinkers and Writers.* New York: Verso, 1998. xv, 206 p. Cloth $27.00.

SUGARMAN, SUSAN: *Freud on the Acropolis: Reflections on a Paradoxical Response to the Real.* Boulder: Westview, 1998. vii, 111 p. Cloth $39.00.

SURBER, JERE PAUL: *Culture and Critique: An Introduction to the Critical Discourses of Cultural Studies.* Boulder: Westview, 1997. ix, 294. Cloth $69.00, paper $24.00.

SUNSTEIN, CASS R.: *One Case at a Time: Judicial Minimalism on the Supreme Court.* Cambridge: Harvard, 1999. xiv, 290 p. Cloth $29.95.

SUTTON, JOHN: *Philosophy and Memory Traces: Descartes to Connectionism.* New York: Cambridge, 1998. xvii, 372 p. Cloth $69.95.

TANESINI, ALESSANDRA: *An Introduction to Feminist Epistemologies.* Malden, MA: Blackwell, 1999. viii, 288 p. Paper $26.95.

TAUBER, ALFRED I.: *Confessions of a Medicine Man: An Essay in Popular Philosophy.* Cambridge: MIT, 1999. xviii, 159 p. Cloth $25.00

TAYLOR, KENNETH: *Truth and Meaning: An Introduction to the Philosophy of Language.* Malden, MA: Blackwell, 1998. xv, 399 p. Paper $27.95.

**Verbal Reasoning Item Writers Needed**

The Association of American Medical Colleges is seeking qualified persons to write verbal reasoning items for a nationally-administered standardized test. Applicants must hold an advanced degree (Ph.D. preferred) in the humanities or social sciences and possess superior writing and reasoning skills.

The work involves a secure test; we ask that a writer's association with it be kept confidential. Applicants should have no recent (five years prior) connection with a commercial test-preparation organization and should be prepared to agree not to establish such a connection in the next five years.

In developing a test unit, writers are asked to follow the criteria described in a writer's guide, which is provided. A test unit consists of a passage of about 550 words, ten multiple-choice questions related to the passage, and a rationale for each response option. New writers are usually allowed one month to complete a trial assignment on a specified topic, for which they receive $450.00.

Prospective item writers should send their vita to:

Mr. Jack Hackett
Association of American Medical Colleges
2501 M Street, NW
Washington, DC 20037

THE JOURNAL OF PHILOSOPHY

VOLUME XCVI, NO. 9, SEPTEMBER 1999

CAMPUS
3-1-2000

## MEASURING CONFIRMATION*

When belief comes to be seen as involving degrees of confidence rather than simple yes-or-no judgments, the theory of rational belief must be adjusted accordingly. Probabilistic coherence stands out as an attractive analogue to the traditional desideratum of deductive consistency. And probability theory, as developed in Bayesian theories of rational belief, seems to go beyond this. It provides a basis for conditionalization principles regulating change of belief—a topic about which traditional logic had little to say; and it offers a quantitative analysis of our notion of confirmation, or evidential support. Confirmation, like belief itself, is something that obviously comes in degrees, but formal confirmation models based on deductive logic are merely qualitative. It seems that the relation these accounts seek to model must derive from a deeper, quantitative one; and Bayesians seem to be in a good position to provide one. According to the standard "positive relevance" account offered by Bayesians, a bit of evidence $E$ confirms hypothesis $H$ to the extent that it makes $H$ more probable, in the sense that $\mathrm{pr}(H/E) > \mathrm{pr}(H)$.

Now, this analysis cannot be thought to match our common-sense concept of evidence exactly. For one thing, the probabilistic relation of positive relevance is symmetrical: $E$ will confirm $H$ on this analysis just in case $H$ confirms $E$. But our ordinary concept of evidence involves a certain kind of epistemic asymmetry. Hypotheses are commonly taken to be supported by evidence, but not vice versa.

---

* I would like to thank Jim Joyce, Mark Kaplan, Hilary Kornblith, Don Loeb, Patrick Maher, Derk Pereboom, and Lyle Zynda for helpful discussions, correspondence, and/or comments on earlier drafts. Thanks also to the University of Vermont for research support.

0022-362X/99/9609/437–61

We do not commonly say, in cases where *E* is evidence for *H*, that *H* is also (at least some) evidence for *E*.

Moreover, our common-sense concept may well be indeterminate, in a way that will necessarily fail to match up with any precise formal model. Some have argued that the standard model, which is based on the *difference* between pr(*H*/*E*) and pr(*H*), should be replaced by a model based on the *ratio* of these quantities, or on the logarithm of this ratio.[1] In support of such suggestions, it has been pointed out that, for instance, a bit of evidence that raised the probability of a hypothesis from $1/10^9$ to 1/100 would intuitively seem much more momentous than one that raised the probability of a hypothesis from .26 to .27.[2] On the ratio-based model, but not on the difference-based model, the former bit of evidence would be seen as more confirmatory. But there are intuitions on the other side here, too. If we compared a bit of evidence that raised the probability of a hypothesis from .0001 to .001, it would likely strike us as much less powerful than a bit of evidence that raised a hypothesis from .1 to .9. This supports the difference model over the ratio model.[3] What is a confirmation theorist to do? One might seek to settle the dispute in favor of one of the two measures by supporting some of these intuitions and/or undermining others. Or one might reject both measures, and try to accommodate all the intuitions in some third measure. But, to my mind, it is not at all clear that this sort of intuitive conflict should be seen as demanding formal solution. One may instead see it as a symptom of a relatively straightforward indeterminateness of the ordinary concept of evidential support.

Compare, for example, the notion of financial support. Suppose we are asked how we should measure the degree to which a candidate *C* is supported by an interest group *I*. One way of measuring this relation would be in terms of the number of dollars *C* received from *I*. But this measure would miss some important aspects of the situation. We might ask what proportion of *C*'s funds are supplied by *I*. Or, looking at the matter from a third angle, we might ask how crucial *I*'s support is for *C* (it might not be crucial even if *I* is a high-dollars, high-percentage contributor, if *C* has sufficient other funds). Or we could use still another measure and ask what *C*'s position would be if $I$ were the only contributor. Thinking about these different measures of support suggests to me that there is no single clear-cut question being asked when we ask 'How much support does $C$ get from $I$?' It would not be surprising if the same were true of the question 'How much does evidence $E$ support hypothesis $H$?'

---

[1] For the former suggestion, see George N Schlesinger, "Measuring Degrees of Confirmation," *Analysis*, LV, 3 (1995). 208-12, for the latter, see Peter Milne, "A Bayesian Defence of Popperian Science?" *Analysis*, LV, 3 (1995): 213-15.

[2] See Schlesinger, pp. 210-11

[3] See Elliott Sober, "No Model, No Inference: A Bayesian Primer on the Grue Problem," in Douglas Stalker, ed., *Grue! The New Riddle of Induction* (Chicago: Open Court, 1994), p. 228, where the example is credited to Ellery Eells.

Nevertheless, despite the fact that it fails to capture epistemic asymmetry, and despite the apparent indeterminateness of our pretheoretic notion, the positive-relevance model seems to be reflecting something important. Surely, 'making more probable' corresponds to something central in our notion of evidential support. The model has been vigorously employed in Bayesian explanations of various tenets of scientific (and common-sense) epistemology. It would, I think, be surprising if some notion along the lines of positive relevance were not at least an important component of our natural notion of evidence.[4]

## I QUANTITATIVE CONFIRMATION AND OLD EVIDENCE

Unfortunately, this model (along with the ratio- and log-of-ratio-based variants) seems to have a fatal flaw. The problem first emerges when considering cases in which $\mathrm{pr}(E)$ is taken to have the extreme value of 1. In those circumstances, $E$ is treated probabilistically just as if it were a tautology. It cannot confirm anything, since $\mathrm{pr}(H/E)$ will be equal to $\mathrm{pr}(H)$. Once one is absolutely certain about one's evidence, it no longer is evidence.[5] Unlike the asymmetry and indeterminateness problems, the problem of old evidence threatens to vitiate the probabilistic approach to confirmation.

Against this version of the problem, it may be protested—reasonably enough, by common Bayesian standards—that we should never believe a nontautology with probability 1 anyway. But this response does not remove the problem's sting. For it remains true that as $\mathrm{pr}(E)$ approaches 1, the degree to which $E$ can confirm anything becomes vanishingly small. The more confident we become in our evidence, the less it can *be* evidence.

This quantitative version of the old-evidence problem is not restricted to the standard relevance measure. It also infects the ratio-

---

[4] What the natural intuitive notion is may not be entirely clear. The core of the intuitive notion I wish to explore here takes confirmation as a relation of support between propositions which explains why confidence in the supporting proposition can help make rational confidence in the supported proposition, where the confidence involved comes in degrees. One might alternatively see confirmation as most fundamentally aimed at explaining rational *acceptance* of propositions. If one took this latter approach, the issues discussed in this paper might look quite different. See, for example, Isaac Levi, *The Enterprise of Knowledge* (Cambridge: MIT, 1980).

[5] Credit for pointing out this problem goes to Clark Glymour, see his *Theory and Evidence* (Princeton: University Press, 1980), pp 85-93

and log-of-ratio-based variants mentioned above. As John Earman points out, the problem also infects Haim Gaifman's confirmation measure (1-pr($H$))/(1-pr($H/E$)).[6] What these measures all have in common is that they ultimately root confirmation in the contrast between pr($H$) and pr($H/E$). As pr($E$) nears 1, this contrast disappears.

A somewhat different probabilistic measure that has been championed by I. J. Good[7] is different in this respect; it is based on contrasting the *likelihoods* pr($E/H$) and pr($E/{\sim}H$). He calls the ratio of these quantities "the (Bayes) factor in favor of $H$ provided by $E$," and the logarithm of this ratio "the weight of evidence in favor of $H$ provided by $E$." The two likelihoods compared in Good's measure do not converge as pr($E$) approaches 1 in quite the same way as pr($H$) and pr($H/E$) do. Nevertheless, this measure, too, turns out to be infected by quantitative old-evidence difficulties. In cases where $E$ confirms $H$, the likelihood ratio ranges between 1 and $\infty$. But when pr($E$) is high, and pr($H$) is moderate—perhaps the paradigmatic situation in which one discusses evidence—the likelihoods are forced to be so close that their ratio falls almost all the way toward the minimum end of this spectrum.[8]

One response to this pervasive problem is simply to give up on finding a Bayesian analysis of our quantitative notion of confirmation. Suppose we discount the cases where pr($E$) = 1, on the grounds that a model of ideal rationality need not account for cases where agents give probability 1 to nontautologies. Once we do this, we are left with an attractive analysis of *qualitative* confirmation, for pr($H/E$) can still be a tiny bit higher than pr($H$) when pr($E$) is very close to 1.[9] But many Bayesians have been reluctant to take this line. A literature has grown up around trying to solve the problem of old evidence.

---

[6] Gaifman, "On Inductive Support and Some Recent Tricks," *Erkenntnis*, XXII (1985). 5-21; and Earman, *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory* (Cambridge: MIT, 1992), p. 121, fn. 8.

[7] See, for example, "Some Logic and History of Hypothesis Testing," in his *Good Thinking: The Foundations of Probability and Its Applications* (Minneapolis: Minnesota UP, 1983). Thanks to Patrick Maher for prompting me to consider Good's measure.

[8] For example, let pr($E$) be .99. If pr($H$) is .75, the highest value the ratio can have is about 1.04. (To see this, note that in general, pr($E$) = pr($H$)pr($E/H$) + pr(${\sim}H$)pr($E/{\sim}H$). So in the present case, we get: .99 = .75pr($E/H$) + .25pr($E/{\sim}H$). From this, it is clear that as pr($E/H$) gets higher, pr($E/{\sim}H$) gets lower, and that the ratio of pr($E/H$) to pr($E/{\sim}H$) will be greatest when pr($E/H$) is as large as possible—that is, when it is 1. Assuming the ratio is thus maximized, we get: .99 = .75 + .25pr($E/{\sim}H$), so pr($E/{\sim}H$) = .96 Thus, the maximum value the likelihood ratio can take when pr($E$) = .99 and pr($H$) = .75 is 1/.96, which is about 1.04.) The value is even lower if we suppose more uncertainty about $H$; if pr($H$) is .6, the ratio maximum drops to under 1.03. And even if we then allow pr($H$) to reach .9, the ratio has a maximum of about 1.1.

[9] For a nice defense of this position, see Mark Kaplan's *Decision Theory as Philosophy* (New York: Cambridge, 1996), chapter 2.

One of the accomplishments of this literature has been to identify different problems of old evidence. Some detailed taxonomies have been offered, but here I chiefly want to distinguish two species of problem. One may think of the problem, as Clark Glymour (*op. cit.*) seems to have, as a problem for the Bayesian theory of change of belief. The problem here is that scientists seem to use evidence of which they have long been confident to *raise* the probability of a hypothesis. This can happen when the hypothesis is thought up at a point when the evidence is already known, or when someone suddenly realizes that the hypothesis entails or explains the evidence. Most of the old-evidence literature is primarily focused on this sort of difficulty for the Bayesian account of belief change; let us call this the *diachronic problem of old evidence.*

The other sort of old-evidence problem does not (at least directly) involve Bayesian accounts of belief change. 'Confirmation', after all, is not just the name of a kind of event. It also seems to refer to a certain relation between propositions. Some propositions seem to help make it rational to believe other propositions. When our current confidence in *E* helps make rational our current confidence in *H*, we say that *E* confirms *H*. For such a relation to obtain, no event consisting of our becoming more confident in *H* on the basis of our confidence in *E* need be occurring.

Indeed, the focus of traditional confirmation theory abstracted completely from particular believings. For the confirmation relation to hold between *E* and *H*, it was not required that *E* or *H* actually be believed at all. The confirmational relations were determined by logical relations; the connection to belief was just that, when *E* confirmed *H*, belief in *E* *would* help make belief in *H* rational.[10] Now, contemporary Bayesians—at least those who have given up Rudolf Carnap's hope of finding the objectively rational probabilistic relations between propositions[11]—cannot so easily separate confirmational relations from relations involving an agent's degrees of belief in propositions. Still, there seems to be a clear sense in which, given the beliefs of a particular agent, certain propositions support others to various degrees. The hope is that these confirmational relations (for a given agent at a given time) will be determined by the agent's degrees of belief (which are, of course, logically constrained by the probability axioms).

---

[10] See, for example, Carl G. Hempel, "Studies in the Logic of Confirmation," in his *Aspects of Scientific Explanation* (New York: Free Press, 1945), p. 24.

[11] One who has not given up on the objective approach is Maher. In "Subjective and Objective Confirmation," *Philosophy of Science*, LXIII (1996): 149-74, he offers a highly contextualized, yet objective, probabilistic account (though of a purely qualitative concept of confirmation).

Unfortunately, analyzing this sort of confirmation in the standard Bayesian way yields the result that my high degree of confidence in $E$ makes $E$ incapable of lending significant rational support to any other belief. Let us call this the *synchronic problem of old evidence.*[12]

## II THE INTEREST OF THE SYNCHRONIC PROBLEM

Various writers have tied the diachronic problem of old evidence to the fact that real agents are imperfect logicians. An agent who is confident of $E$ may fail to see that it follows from $H$, and then, upon discovering the entailment, raise her confidence in $H$. This has led to one of the two main lines of approach to the old-evidence problem. The standard notion of probabilistic coherence requires agents to have a certain sort of logical omniscience: they must give probability 1 to all logical truths, and respect all logical entailments perfectly (for example, by never having $\mathrm{pr}(P) > \mathrm{pr}(Q)$ when $P$ entails $Q$). Once this requirement is relaxed, we might devise ways of describing how learning *that H entails E* can—via probabilistic calculation—raise an agent's confidence in $H$.[13]

This approach treats the old-evidence problem as a symptom of excessive idealization. Clearly, even the most epistemically perfect among us cannot approach logical omniscience. A theory that builds such omniscience centrally into its epistemic model may lose the ability to give a rational reconstruction of much human behavior that is surely epistemically praiseworthy. Thus, the project of making the Bayesian theory of rationality more realistic is clearly an important one.

---

[12] My distinction between the diachronic and synchronic problems mirrors Daniel Garber's distinction between the *historical* and *ahistorical* problems, and Lyle Zynda's distinction between the problem of the *confirmation event* and that of the *confirmation relation.* It more roughly corresponds to Eells's distinction between the problem of *new old evidence,* and the problems of *old new evidence* and *old old evidence.* See Garber, "Old Evidence and Logical Omniscience in Bayesian Confirmation Theory," in Earman, ed., *Testing Scientific Theories, Minnesota Studies in the Philosophy of Science,* Volume X (Minneapolis: Minnesota UP, 1983), pp 99-131; Zynda, "Old Evidence and New Theories," *Philosophical Studies,* LXXVII (1995): 67-95; and Eells, "Problems of Old Evidence," *Pacific Philosophical Quarterly,* LXVI (1985): 283-302.

[13] See Garber; Richard Jeffrey, "Bayesianism with a Human Face," in *Testing Scientific Theories;* and Ilkka Niiniluoto, "Novel Facts and Bayesianism," *British Journal for the Philosophy of Science,* XXXIV (1983) 375-79.

Nevertheless, there is a sense in which the diachronic old-evidence problem, as a symptom of overidealization, leaves an important aspect of the Bayesian model unscathed. It does not show Bayesianism defective as a certain kind of epistemic ideal, or model of rational perfection. It is quite plausible that a being who had unlimited intellectual powers, but limited data, *would* have a probabilistically coherent confidence distribution over all possible hypotheses. For such an ideal being, the diachronic problem of old evidence simply would not arise. This observation at least limits the urgency of the diachronic old-evidence problem, if one is willing to allow for a good deal of idealization in thinking about epistemic norms. And in fact, the degree of idealization involved here should not be too shocking for those accustomed to thinking about ideally rational all-or-nothing beliefs in terms of the consistency and closure conditions of deductive logic.[14]

But the bite of the old-evidence problem goes deeper than revealing the extent of idealization inherent in Bayesianism. This comes out in considering the synchronic version of the problem. Consider an agent who is not only currently logically omniscient, but who has always been so, and who has arrived at her present beliefs by faultless application of the idealized mechanisms of Bayesian belief revision. Even for such an ideal intellect, it seems clear that two things can be true simultaneously: (1) some beliefs will provide significant evidential support for others; and (2) some of the evidential beliefs will be held with high levels of confidence. These two features are, as noted above, flatly incompatible on the standard Bayesian analysis of (synchronic) evidential support. Thus, no lapse from rational perfection, even in the past, is needed for the synchronic old-evidence problem to arise. For this reason, the synchronic problem presents a particularly sharp challenge to the philosophical interest of the Bayesian program: it shows standard Bayesianism defective even as a model of idealized rationality. In what follows, then, I shall focus on the synchronic problem.

---

[14] It might be objected that one who was comfortable with a lot of idealization would have little reason to prefer Bayesianism over traditional logic in the first place. Is not probability theory itself basically a response to excessive idealization of the all-or-nothing model of belief?

Such a comparison would be misleading. The problem with the bivalent model of belief is not just that it fails to describe actual imperfect human attitudes toward propositions Even a rationally ideal being would not have all-or-nothing beliefs on contingent matters about which she was incompletely informed. An agent with this limited information who nevertheless harbored only bivalent beliefs would be less epistemically perfect, not just less realistic, than one whose confidence in such propositions came in degrees. Since having only all-or-nothing beliefs based on incomplete information is not a rational ideal, the move to Bayesianism is not one of trading off idealization for realism, but one of improving the idealization.

## III. CAN THE SYNCHRONIC PROBLEM BE REDUCED TO THE DIACHRONIC?

One reason that interest has centered around the diachronic problem may be based on the thought that a solution to the diachronic problem would provide a solution to the synchronic one. Suppose that we had in hand a satisfactory account of confirmation events—events in which probabilities of hypotheses are raised by evidence. A given agent's body of belief will be the result, in part, of a history of such events. Given such a history, Ellery Eells[15] writes that "it seems appropriate to say that *E is (actual) evidence for T*, for a given individual, if, at some time in the past, the event of its confirming *T*, for that individual, took place" (*op. cit.*, p. 287). Earman[16] concurs with this approach, which I shall call the *historical* approach to the synchronic problem.

The historical approach clearly gives up on the idea that confirmational relations (for a given agent at a given time) are determined by the agent's present degrees of belief. Still, there is something attractive about the idea. And if it were successful, the synchronic problem of old evidence would be reduced to the diachronic problem. In that case, the entire old-evidence problem might seem to stem from over-idealization.

It turns out, however, that the historical approach overestimates the connection between the two concepts of confirmation. To see this, consider an agent who is wondering whether deer live in a nearby wood. He comes across a pile of deer droppings, and his confidence in the deer hypothesis increases to near 1. Shortly thereafter, he finds a shed antler. Since his confidence in the deer hypothesis is already so high, this new evidence does not have any significant impact on it. Now, subsequent to the agent's finding the second piece of evidence, let us ask whether the evidence about the droppings and the antler provide rational support for our agent's belief about deer. Intuitively, the droppings and the antler provide equally strong rational support for the agent's deer belief. There is, I think, no sense in which the droppings currently provide stronger support or better evidence than does the antler. But only the droppings are historically associated with a significant increase in the agent's probability for deer. The historical approach thus makes contemporary evidential support depend in an unintuitive way on the order in which evidence was discovered.[17] Intuitively, synchronic support should depend on the agent's present epistemic state, not on such historical accidents. Clearly, the synchronic problem cannot be reduced to the diachronic one in this way.[18]

---

[15] Eells distinguishes 'confirming' from 'being evidence for', but I shall use the terms interchangeably. His analysis is intended to capture the sense in which *E* supports *H* after *E* is known.

[16] *Bayes or Bust?*, chapter 5.

## IV. THE COUNTERFACTUAL APPROACH

The most immediately appealing, and most widely criticized, direct approach to the synchronic problem relies on counterfactual degrees of belief. For an agent who knows *E*, this approach would analyze confirmation with the standard positive-relevance definition; however, it would be applied not to the agent's actual probabilities, but to what those probabilities would have been in circumstances where the agent did not know *E*.[19]

This approach, like the historical approach, would abandon seeing confirmation relations as determined by an agent's probabilities. Still, the intention seems to be to make confirmation relative to a probability function that is in a sense as close to the agent's as possible, consistent with the agent's not knowing *E*.

---

[17] I make this argument briefly in a review of Earman's *Bayes or Bust?* in *Philosophical Review*, CIII (1994): 345-47. Maher argues for the same conclusion using a somewhat different type of example. In his example, the bearing of the evidence on a hypothesis changes as background knowledge changes. In such situations, a bit of evidence that intuitively confirms a certain hypothesis today might have disconfirmed it when it was first discovered.

[18] One might object that this sort of problem need not trouble a historical analysis of *qualitative* confirmation. (Earman is clearly interested in a Bayesian analysis of the quantitative notion, but Eells may have intended his argument to apply only to a qualitative notion.) The point would be that finding the antler presumably would have raised the probability of deer at least a tiny bit, given that the probability for deer would always remain a bit shy of unity.

But as mentioned above, it is not obvious that the old-evidence problem arises at all, if we seek only a qualitative analysis of confirmation and preclude complete confidence in contingent matters. In this case, it is not clear that there is any reason to substitute the historical account for the standard, nonhistorical analysis.

On the other hand, if we do allow agents to believe contingent matters with probability 1, then it would be hard to see why the deer hypothesis could not attain this level. And if it did, then the historical account of even qualitative confirmation would fail in just the manner described in the text. Evidence that was discovered subsequent to the hypothesis' being given probability 1 could not increase its probability. Such evidence might nevertheless provide just as much rational support for the hypothesis as would the evidence that historically helped raise its probability.

Finally, the sort of example given by Maher would tell against qualitative versions of the historical account even if we did not allow full belief in contingent propositions.

[19] This approach is tentatively endorsed in Garber and embraced more wholeheartedly in Colin Howson, "Bayesianism and Support by Novel Facts," *British Journal for the Philosophy of Science*, XXXV (1984): 245-51.

But will the counterfactual move allow us to capture what we want to capture from the agent's actual probability function? Various writers[20] have noticed problems here. For one, it is not clear that the counterfactual stipulation determines any unique probability function. Moreover, in certain cases, the distribution that the agent would have in the counterfactual situation would be produced by irrational factors that would accompany the nonbelief in $E2$; in such cases, the counterfactual degrees of belief give the wrong result.[21] These sorts of cases highlight the problem of getting too far away from the agent's real probabilities that were supposed to ground the confirmation relation. The factors that supplement the agent's probabilities in determining the truth of the counterfactuals are often irrelevant to the agent's epistemic state.

## IV. DOES JEFFREY CONDITIONING UNDERMINE THE SYNCHRONIC NOTION?

Another approach to the synchronic problem is to argue that, from the Bayesian point of view, the whole project of analyzing synchronic confirmation should be abandoned. Lyle Zynda (*op. cit.*) argues that, once we see the Bayesian position as allowing belief updating by Jeffrey conditioning as well as classical conditionalization, we can see that the whole Bayesian project of analyzing synchronic confirmation falls apart—and for reasons other than the old-evidence problem.[22]

The argument begins by considering cases where $\mathrm{pr}(E)$ is moderate, so the old-evidence problem does not arise. For such cases, Zynda suggests that we should see $E$ as confirming $H$ only if the following condition holds: for any probability function PR that comes from the present probability function pr by Jeffrey conditioning initiated by a set of probability changes which includes raising the probability of $E$, $\mathrm{PR}(H) > \mathrm{pr}(H)$.

Zynda shows that this condition is not acceptable. An instance of Jeffrey conditioning may be initiated by changes in the probabilities of several propositions. Even when $E$ is intuitively confirmatory, an increase in $E$'s probability may be accompanied by decrease in probability of an even-more-strongly confirmatory proposition $F$, resulting in a decrease of probability for $H$ overall. Zynda considers modifying the condition to specify that no other positively relevant proposition is decreased in probability; however, he shows by example that even this modified condition excludes some cases in which *E* intuitively confirms *H*.

---

[20] See, for example, Glymour; Eells; Charles S. Chihara "Some Problems for Bayesian Confirmation Theory," *British Journal for the Philosophy of Science*, XXXVIII (1987): 551-60; and Earman, *Bayes or Bust?*

[21] See Maher.

[22] The classical conditionalization principle changes a probability function in response to an evidential proposition's being raised to 1. Jeffrey conditioning changes a probability function in response to initiating changes in the probabilities of one or more propositions. These initiating changes may be in either direction, and may be to nonextreme probability values.

Zynda concludes that "Jeffrey conditioning introduces a new, quite radical kind of contextualism: the evidential value a given proposition has cannot be evaluated independently of what is happening in other regions of logical space" (*op. cit.*, p. 77).

> Judgments of evidential support have to be relativized to particular ways of updating, i.e. to particular epistemic events. Once we have a particular epistemic event in mind...judgments of evidential support make sense.... The quite different project of coming up with an analysis of "*e* confirms *h* for *S* at *t*" is not in and of itself a very fruitful one for Bayesians to pursue (*op. cit.*, p. 79).

If this argument is correct, we should give up the project of analyzing synchronic confirmation relations altogether. Once we admit Jeffrey conditioning, the evidential value of propositions is contextualized in a way that precludes our meaningfully asking in general whether *E* confirms *H* for *S* at *t*.

It seems to me, however, that the project of analyzing the synchronic notion of confirmation should not be dismissed this quickly. Recall that, in response to the first difficulty, Zynda modified his condition to exclude cases involving a certain kind of interference—cases where the probability of some other positively relevant proposition *F* was reduced while *E*'s probability was raised. But the example to show that the modified condition was still unacceptable involved the interfering effects of another evidential proposition—a negatively relevant proposition whose probability was raised. This suggests a further modification, in which the condition would also specify that this second sort of interference not occur. The basic idea here would be that, in Jeffrey conditioning, raising the probability of a confirming proposition *E* will, ceteris paribus, raise the probability of *H*. The ceteris paribus condition should naturally preclude both simultaneous reduction of probability for other confirming propositions, and simultaneous increase in probability for disconfirming propositions.

Thus remodified, the condition avoids the difficulties that faced its predecessors. So the fact that Jeffrey conditioning can involve multiple evidential propositions simultaneously does not, after all, render the evidential effect of any one of these propositions contextualized in a way that would prevent us from asking meaningfully whether a given evidential proposition confirms or disconfirms a given hypothesis (for a given agent at a given time). Indeed, it seems that allowing Jeffrey conditioning leaves us in much the same position with respect to the synchronic notion of confirmation.[23]

## VI. A SOLUTION TO THE SYNCHRONIC PROBLEM?

Why does increasing pr($E$) drain $E$ of confirmatory power, as measured by the standard account? The standard account measures confirmation by how much increasing pr($E$) to 1 would raise pr($H$). This partly realizes the traditional confirmation-theoretic idea that $E$'s confirming $H$ does not require $E$ to be believed. But, curiously, it does not fully accord with the idea that our current degree of belief in $E$ is irrelevant to the question of how much $E$ confirms or disconfirms $H$. Consider a situation in which an agent's beliefs change by her becoming increasingly confident in $E$. As pr($E$) increases, pr($H$) increases to take account of the increased confidence in $E$. Thus, the probability of $H$ on the assumption that $E$ is true, and the probability of $H$ all things considered, become closer. In a sense, $E$ just becomes part of "all things considered." At this point, becoming more certain of $E$ cannot raise the probability of $H$ much more, since there just is not room to become much more confident in $E$. In such circumstances, the standard measure shows very little evidential support. Yet intuitively, the agent's confidence in $E$ is already an important support for her present confidence in $H$. This dimension—we might call it the dimension of *actualized support*, as opposed to *potential further support*—is simply ignored by the standard measure. This is the problem of old evidence.

From this perspective, we can also see why the traditional measure suffers from what might be called the *probable-hypotheses problem*. This is a difficulty that is very similar to the old-evidence problem, but much less celebrated.[24] Consider cases where pr($H$) gets very high. Again, there is in such cases little room for $E$ (any $E$) to raise pr($H$) further. The traditional measure, taking into account only potential further support, reflects this fact; once pr($H$) is close to 1, no evidence will be counted as providing a significant level of support for *H*. Oddly, once we become quite confident in a hypothesis, we are counted as having no significant evidential basis for our confidence!

---

[23] This impression is reinforced by the observation that the remodified condition will be satisfied whenever the standard positive-relevance condition is satisfied. My hunch to this effect was proved by Zynda (in correspondence; I should also note that he expressed his current reservations about the strong claim criticized above). A related theorem is proved by Carl G. Wagner in "Old Evidence and New Explanation," *Philosophy of Science*, LXIV (1997): 677-91.

[24] This problem—or at least the qualitative version that occurs when pr($H$) = 1—is briefly celebrated by Kaplan (p. 50), who notes its similarity to the old-evidence problem.

It seems crucial, then, that our measure of confirmation go beyond measuring potential further support. The question then arises whether a dimension of actualized support is somehow reflected in the agent's current probability function. Intuitively, the standard measure reflects the amount that pr($H$) would rationally increase just as a result of the agent raising her pr($E$) to 1. This potential increase in pr($H$) might be thought of as resulting from two factors: the degree to which $E$ is linked to $H$, and the distance $E$ has to travel in attaining probability 1. What we are interested in is the former factor, not the latter.

How might we control for the latter factors' influence? The most obvious way would be to normalize in a way that corrects for it. For example, if pr($E$) is already .75, so that it can be raised by only 1/4 of the distance between 0 and 1, we can multiply the potential increase in pr($H$) by 4. If pr($E$) is already .99, so that it can be raised by only 1/100 of the distance between 0 and 1, we would multiply the potential increase in pr($H$) by 100. It turns out that dividing the standard measure by pr($\sim E$) achieves just this result. This suggests using the following as a measure of confirmational support:

$$S(E,H) = [\mathrm{pr}(H/E) - \mathrm{pr}(H)]/\mathrm{pr}(\sim E)$$

Now, one might immediately object to this measure that, when pr($E$)=1, it is undefined. But as noted above, there are good reasons to hold that contingent propositions should never be given probability 1. Thus, we may stipulate that the definition—which is, after all, part of an account of ideal reasoning—apply only when $0<\mathrm{pr}(E)<1$. This tack is especially appealing if one is trying to analyze confirmation in terms of standard probabilistic relations among an agent's beliefs, since from this point of view, propositions having probability 1 are indistinguishable from tautologies. This is important because it strongly suggests that on a standard probability-based approach, it would be hopeless to try to measure some interesting kind of evidential support from such propositions.

Interestingly, there is another intuitive route to essentially the same measure. We want to capture the support an agent's confidence in *H* *already* receives from *E* (in contrast to the potential further support that might be gotten from raising pr($E$)). We might measure this by comparing the agent's present probability for $H$—which includes that present support—with her probability for $H$ *on the condition that $E$ is not true*: $\mathrm{pr}(H) - \mathrm{pr}(H/{\sim}E)$. Adding this dimension of actualized support to the traditional measure of potential further support yields:

$$S^*(E,H) = \mathrm{pr}(H/E) - \mathrm{pr}(H/{\sim}E)$$

as a measure combining both sorts of support. If conditional probabilities are defined in the standard way, $S^*$ is also undefined when $\mathrm{pr}(E)=1$. And in cases where $0<\mathrm{pr}(E)<1$, measure $S^*$ turns out to be equivalent to measure $S$.[25]

$S^*$ is briefly considered, and rejected (for reasons that will be examined below) by Earman.[26] His reason for considering it is not either of the intuitive arguments given above. Rather, he points out that conditional probabilities can be defined in a (nonstandard) way which allows $\mathrm{pr}(H/{\sim}E)$ to be defined when $\mathrm{pr}(E)=1$. If we do this, then $S^*$, unlike the standard positive-relevance measure, can take positive values when $\mathrm{pr}(E)=1$. But the measure has other advantages, in addition to the intuitive naturalness explained above, which strike me as more important. Since the remaining discussion will not turn on interpreting conditional probabilities nonstandardly, I shall treat $S^*$ as an alternate formulation of $S$, and

[25]

$$S(E, H) = \frac{\mathrm{pr}(H/E) - \mathrm{pr}(H)}{\mathrm{pr}({\sim}E)}$$

Expanding $\mathrm{pr}(H)$, we get:

$$\frac{\mathrm{pr}(H/E) - [\mathrm{pr}(H/E)\mathrm{pr}(E) + \mathrm{pr}(H/{\sim}E)\mathrm{pr}({\sim}E)]}{\mathrm{pr}({\sim}E)}$$

Regrouping the numerator gives:

$$\frac{[\mathrm{pr}(H/E) - \mathrm{pr}(H/E)\mathrm{pr}(E)] - \mathrm{pr}(H/{\sim}E)\mathrm{pr}({\sim}E)}{\mathrm{pr}({\sim}E)}$$

Factoring $\mathrm{pr}(H/E)$ out of the left side of the numerator yields:

$$\frac{\mathrm{pr}(H/E)[1 - \mathrm{pr}(E)] - \mathrm{pr}(H/{\sim}E)\mathrm{pr}({\sim}E)}{\mathrm{pr}({\sim}E)}$$

Since $[1 - \mathrm{pr}(E)] = \mathrm{pr}({\sim}E)$, this may be canceled out, leaving:

$$\mathrm{pr}(H/E) - \mathrm{pr}(H/{\sim}E) = S^*(E, H)$$

[26] *Bayes or Bust?*, pp. 120-21.

shall use whichever formulation is best suited to the purpose at hand.[27]

As Earman notes, $S$ agrees qualitatively with the standard measure in cases involving nonextreme probabilities. (This can be seen most easily by considering the first formulation given above. Since $S$ simply divides the standard measure by a positive number ($\mathrm{pr}(\sim E)$), positive (negative) confirmation will remain positive (negative).) Given that the standard measure is qualitatively plausible in cases involving nonextreme probabilities, this is an advantage.

More impressive, however, is $S$'s quantitative treatment of old evidence. The most troubling version of the old-evidence problem occurs when $\mathrm{pr}(E)$ is high, but still short of unity. Unlike the standard measure, $S$ allows *significant* confirmation even when $\mathrm{pr}(E)$ is, say, .999999. This is an enormous advantage. It also suggests that $S$ may come closer than the standard measure to satisfying the intuition that the degree to which $E$ confirms $H$ should not depend on how confident we are of $E$.

In fact, $S$ is, in an intuitively appealing way, insensitive to $\mathrm{pr}(E)$. Suppose an agent has a probability function relative to which the support $E$ gives to $H$ as measured by $S$ (let us call this $S$-*support*) has some particular value. Let us change the agent's probabilities by Jeffrey conditioning initiated by raising (or lowering) just the probabilities for $E$ and $\sim E$. It would be nice if, relative to the agent's new probability function, the $S$-support $E$ gives to $H$ retained the same value. In fact, this is the case: in general, $S$-support given by $E$ is stable over Jeffrey conditioning on $\{E, \sim E\}$.[28]

As an illustration, consider a rational agent confronted by a standard die he believes to be fair, which has been rolled, but is out of sight. Let $O$ be the proposition that an odd number came up, and $L$ be the proposition that a low number (1-3) came up. For such an agent, $\mathrm{pr}(O)=1/2$ and $\mathrm{pr}(L)=1/2$. Using the second formulation given above, $S(L,O) = \mathrm{pr}(O/L) - \mathrm{pr}(O/{\sim}L) = 2/3 - 1/3 = 1/3$: a modest degree of positive support.

---

[27] As this paper was under consideration, Jim Joyce sent me a manuscript chapter of his forthcoming book, in which he supports a version of measure $S^*$ as capturing one concept of evidential relevance. Joyce would apply the measure even in cases where $\mathrm{pr}(E)=1$, by defining it via Reyni-Popper measures, which work much like standard conditional probabilities, but define $\mathrm{pr}(H/\sim E)$ even in cases where $\mathrm{pr}(\sim E)=0$. Joyce interprets such measures as describing the epistemic state of someone who is fully certain that $E$ is true, but still has opinions as to the probability of $H$ if she is wrong about $E$. Joyce's discussion is highly recommended. While it has some parallels with the present discussion, it has interesting differences and reaches different conclusions; for example, he sees the standard confirmation measure as capturing a different, but no less legitimate, concept of evidential relevance. See Joyce's *The Foundations of Causal Decision Theory* (New York: Cambridge, 1999), chapter 6.

[28] This is easy to see in the $S^*$ formulation. The conditional probabilities by which the measure is defined are unchanged in Jeffrey conditioning, since Jeffrey conditioning involves rigidity of all probabilities conditional on the members of the relevant partition.

Now suppose the agent raises the probability of $L$ to .99 by Jeffrey conditioning in which $L$ and ${\sim}L$ are the only propositions directly affected by experience. Her new probability PR for $O$ will, of course, be significantly higher:

$$\mathrm{PR}(O) = .99\mathrm{pr}(O/L) + .01\mathrm{pr}(O/{\sim}L) = .66 + .00333... = .66333...$$

This higher probability for $O$ intuitively incorporates the agent's increased confidence in $L$. It also renders support as standardly measured almost nonexistent: $\mathrm{PR}(O/L) - \mathrm{PR}(O) = .00333....$ This is a prime example of the synchronic old-evidence problem. The low support figure represents the potential for becoming more certain of $O$ on the basis of becoming even more certain of $L$, but leaves out the extent to which $\mathrm{PR}(O)$ already incorporates support from confidence in $L$.

By contrast, $S$-support is unchanged at 1/3. Thinking in terms of the first formulation, it is easy to see how this figure represents the standard measure of support normalized by being multiplied by 100, since the agent's move to full belief in $L$ would move $\mathrm{PR}(L)$ 1/100 of the way from 0 to 1. Intuitively, it seems that $S$ counts both potential and actualized support. Cases like this suggest that $S$ offers a possible solution to the synchronic problem of old evidence.[29]

## VII. THE OLD-EVIDENCE PROBLEM RETURNS?

As noted above, Earman quickly rejects $S^*$ as a suitable measure of confirmation. His reason is that, in cases where $H$ entails $E$ and

---

[29] It may be of interest to see how this sort of situation is handled by the likelihood-ratio-based model mentioned above. The model would measure confirmation in the dice case by contrasting the likelihoods $\mathrm{pr}(L/O)$ and $\mathrm{pr}(L/{\sim}O)$. In the initial situation, these quantities are 2/3 and 1/3 respectively, so the likelihood ratio is 2. But after the probability of $L$ is raised to .99, the likelihoods take on new values. In particular (using values from the text).

$$\mathrm{PR}(L/O) = \frac{\mathrm{PR}(L)\cdot\mathrm{PR}(O/L)}{\mathrm{PR}(O)} = \frac{.99\cdot(2/3)}{.66333\ldots} = \frac{.66}{.66333\ldots} \approx .99$$

$$\mathrm{PR}(L/{\sim}O) = \frac{\mathrm{PR}(L)\cdot\mathrm{PR}({\sim}O/L)}{\mathrm{PR}({\sim}O)} = \frac{.99\cdot(1/3)}{.33666\ldots} = \frac{.33}{.33666\ldots} \approx .98$$

Thus the likelihood ratio shrinks to about $.99/.98 \approx 1\ 01$.

$\mathrm{pr}(E)=1$, $S^*(E,H)$ is just equal to $\mathrm{pr}(H)$. Thus, for any hypothesis $H$, all evidence which is both (1) entailed by $H$ and (2) believed with certainty is counted as confirming $H$ to the same degree.

Now, this problem as posed by Earman might not seem to concern us if we intend our measure to be restricted to cases where $\mathrm{pr}(E)<1$. But there is a closely related phenomenon that affects even these cases. It derives in part from the fact that, in cases where $H \vdash E$, $S(E,H) = \mathrm{pr}(H/E)$ (this is because $\mathrm{pr}(H/{\sim}E) = 0$). This fact alone does not make all entailed evidence equal, however. For example, consider our agent who has ordinary expectations about a fair unseen die, and who thus has the standard simple and conditional probabilities for the following propositions:

$E2$: the die shows a multiple of two.
$E3$: the die shows a multiple of three.
$H6$: the die shows six.

Both $E2$ and $E3$ are entailed by $H6$. In this situation, both should confirm $H6$, but $E3$ should confirm it more strongly. And measure $S$ gives just this result: $S(E2,H6) = \mathrm{pr}(H6/E2) = 1/3$, while $S(E3,H6) = \mathrm{pr}(H6/E3) = 1/2$.

So far, we have not seen the difficulty. But, as we know from the old-evidence problem, as $\mathrm{pr}(E)$ approaches 1, $\mathrm{pr}(H/E)$ approaches $\mathrm{pr}(H)$. Thus, if the above agent's probabilities for both $E2$ and $E3$ are raised to nearly 1, the probability of $H6$ conditional on each of these propositions will be very close to the unconditional probability of $H6$, and hence the support the two propositions give $H6$ will, according to $S$, be very nearly the same. This is a disturbing result, for intuitively, $E3$ should give more confirmation than $E2$, even when both are highly probable. Does this not show that $S$ encounters the old-evidence problem after all, at least for cases where $H \vdash E$?

Closer inspection reveals, I think, that the answer to this question is "no." This is not the problem of old evidence. To see this, let us begin by considering what might seem an inconsistency between the claims made above about $S$.

Immediately above, we saw that, in a situation involving ordinary beliefs about a hidden die, $S$ assigns $E2$ and $E3$ robustly different degrees of support for $H6$, while if $E2$ and $E3$ both become very highly probable, $S$ will have to assign them very nearly the same degree of support for $H6$. In the previous section, however, we saw that $S$ is, in a very natural sense, insensitive to $\mathrm{pr}(E)$. If this is true, how could raising $\mathrm{pr}(E2)$ and $\mathrm{pr}(E3)$ erase the difference in the $S$-support they give $H6$?

The first thing to notice is that the previous section's claim does hold in the present example. If we begin with ordinary beliefs about a hidden die, and increase belief in, for example, $E2$ by Jeffrey updating on just $E2$ and its negation, $S(E2,H6)$ does not change, even when $\mathrm{pr}(E2)$ gets very high. The probabilities in such a situation will look like this (using PR for the post-updating beliefs, and pr for the initial beliefs):

$$\mathrm{PR}(E2) = .99$$
$$S(E2,H6) = \mathrm{PR}(H6/E2) = \mathrm{pr}(H6/E2) = 1/3$$

This shows that $S(E2,H6)$ remains unchanged. But other things have changed. For one thing, $\mathrm{PR}(H6)$ now reflects the positive evidence from $E2$:

$$\mathrm{PR}(H6) = .99\mathrm{pr}(H6/E2) + .01\mathrm{pr}(H6/{\sim}E2) = .99(1/3) + 0 = .33$$

More interestingly, some probabilities conditional on $E3$ have changed. The new probability for $E3$ itself is still 1/3, since $E2$ is probabilistically irrelevant to $E3$. But:

$$\mathrm{PR}(H6/E3) = \mathrm{PR}(H6 \ \& \ E3)/\mathrm{PR}(E3) = \mathrm{PR}(H6)/\mathrm{PR}(E3) = .33/(1/3) = .99$$

Because of this, the support $E3$ gives to $H6$, according to measure $S$, is now close to maximal:

$$S(E3,H6) = \mathrm{PR}(H6/E3) = .99$$

So, while raising the probability of $E2$ did not change *its* confirmatory power for $H6$, it dramatically increased the confirmatory power of $E3$. Raising the probability of one evidential proposition can thus affect confirmation relations involving other propositions. So the fact that $S(E, H)$ is insensitive to simply raising the probability of $E$ is quite consistent with the fact that raising the probability of two or more evidential propositions will disturb the $S$-confirmation relations involving all of them.

This example also shows that the disturbing fact—that raising the probabilities of both $E2$ and $E3$ would erase the differences in confirmatory power between them—is not simply an instance of the old-evidence problem. For either one of these same propositions can be believed with near certainty without any effect on its confirmatory power. Still, while measure $S$ thus seems to *avoid* the old-evidence problem, it would be premature to conclude that it *solves* the problem. Presumably, a solution to the problem would require a measure which not only avoided it, but which was otherwise reasonable. How unreasonable are the verdicts $S$ delivers in the disturbing cases?

## VIII A DIFFERENT PROBLEM FOR BAYESIANISM

Let us begin by looking more carefully at the problematic case mentioned above. Suppose that an agent starts with ordinary beliefs about a hidden die, and is told first by one highly trusted source that it shows a multiple of two ($E2$), and then by another that it shows a multiple of three ($E3$). Intuitively, $E3$ would seem more important than $E2$ in supporting her now-high confidence that ($H6$) the die is showing a six. But measure $S$ must now give them almost equal confirmatory power. In particular, the confirmatory power of both $E2$ and $E3$ is nearly maximal in this case.[*] Raising the probability of $E2$ raised the confirmatory power of $E3$, and vice versa, to the extent that the difference between them was nearly erased.

One important question to be asked about this example is about exactly where the counterintuitiveness lies. The fact that $E2$ and $E3$ have nearly equal confirmatory power derives from another fact: that each of them has confirmatory power close to the maximum of 1. But this seems to be a problematic fact about each one of the evidential propositions, independent of questions about *differences* in confirmatory power. (If it was intuitively reasonable that raising $\mathrm{pr}(E2)$ would make $E3$ nearly maximally confirmatory, and vice versa, then presumably it would be reasonable for them to have approximately equal confirmatory power in that situation.) If this is right, then the real root of the problem should reveal itself midway through the example story, when the agent comes to believe strongly in $E2$, rendering $E3$ almost maximally confirmatory. What, intuitively, is going on at this point?

Recall that the way learning $E2$ raises the confirmatory power of $E3$ so dramatically is something like this: once $\mathrm{pr}(E2)$ becomes close to 1, $\mathrm{pr}(H6/E3)$ gets close to 1. Since $S(E3,H6)$ is in this case equivalent to $\mathrm{pr}(H6/E3)$, it also rises to nearly 1. Or, more informally, what happens is that when it is almost certain that a multiple of two is showing, the information that a multiple of three is showing is virtually equivalent to the information that a six is showing. Our confirmation measure reflects the strength of this connection.

Why should this sharp increase in confirmatory power be counterintuitive? I think it is useful to compare this case to another. Consider a second agent, who starts off with beliefs appropriate to a fair die having only three (curved) sides, bearing a two, a four, and a six, respectively. For such an agent, it would intuitively be entirely reasonable to take $E3$ as maximally confirmatory. After all, with these background beliefs, $E3$ is virtually equivalent to $H6$. But now let us ask what the difference is between our two agents. From the point of view of their probabilities, the answer seems to be "not much!" For both, pr($E2$) would be nearly 1, pr($E3$) and pr($H6$) close to 1/3, and pr($H6/E3$) nearly 1. For both, learning $E3$ would virtually guarantee $H6$. Why, then, is there such a difference in our intuitions about confirmatory power in these two cases?

---

[*] Consider $E2$: $\mathrm{pr}(H6/E2)$ is almost 1, since it is higher than $\mathrm{pr}(H6)$, which itself is near 1. And $\mathrm{pr}(H6/{\sim}E2)$ is 0, because $H6 \vdash E2$. Thus $S(E2,H6)$ is almost 1.

The answer, it seems to me, is related to the following intuition: in the case with the ordinary die, if $E3$ is learned, it will raise pr($H6$) to near 1, but it will be doing only part of the work of raising it from its baseline level (1/6), the rest being done by $E2$. In the three-sided die case, if $E3$ is learned, it will be doing all of the work in raising the probability of $H6$ from its different baseline level (1/3) to near certainty.

When we think about the evidential power of $E3$ in the case with the ordinary die, we seem to assess it against $H6$'s intuitive base-line level. We intuitively take the elimination of the odd-numbered possibilities as something that was accomplished by another piece of evidence. When we think about the three-sided die case, by contrast, we take the absence of odd-numbered possibilities as being simply part of the given background conditions for our reasoning. This suggests that the intuitive difference between the cases depends on a factor to which our confirmation measure is insensitive: the distinction between specific evidence other than $E$, and background assumptions.

This point constitutes a further departure from the worry expressed by Earman. He rejected a version of measure $S$ because it equalized confirmatory power for all evidence that was both entailed by the hypothesis and had probability 1. We have already seen that essentially the same difficulty occurs at evidential probabilities below 1. Moreover, in one such case, it seems that unintuitive equalization per se is not the basic problem—rather, it is a symptom of unintuitiveness in measure $S$'s verdicts on individual items of evidence. Now, if the essential problem is not about equalizing differences in these cases, but rather about a very general feature of confirmation—the background/evidence distinction—then we might suspect that the problem should occur even in situations where $H$ does not entail $E$.

It seems to me that the fundamental problem is, in fact, quite independent of $H$ entailing $E$. To see this, consider the following variant of our deer example: the agent is wondering whether deer live in a certain wood that he has not yet explored. Let $H$ be 'Deer live in this wood', *D* be 'There are deer droppings at location *x*', and *A* be 'There is a shed deer antler at location *y*', where *x* and *y* are some narrowly circumscribed locations in the wood. This case is in a sense diametrically opposite from the dice case, in that the evidential propositions are not even "close to" being entailed by *H*. (Since *D* and *A* specify particular locations for the droppings and antler, their probabilities even conditional on *H* are not very high.)

Now suppose the agent starts with $\mathrm{pr}(H)=.5$, $\mathrm{pr}(D)=.001$, and $\mathrm{pr}(A)=.0001$. Both *D* and *A* will confirm *H* fairly strongly—at about .5.[21] Suppose, however, that the agent then finds what are almost certainly deer droppings at location *x*, sending the probabilities for *D*, and hence *H*, very high. This does not, of course, change *D*'s level of *S*-support for *H*. This is good news. In fact, it not only shows, like the previous example, that *S* allows old evidence to confirm hypotheses strongly. It also shows that *S* allows instances where *E* strongly confirms *H* even when $\mathrm{pr}(H)$ is very high. Thus, *S* does not seem to suffer from what we called above the problem of probable hypotheses—the inability of the standard measure to assign significant degrees of confirmation to highly probable hypotheses.

Unfortunately, not all the news is good. In particular, although *S*-support from *D* is unchanged, *S*-support from *A* suddenly becomes very low. This is because *H* is now very likely irrespective of *A*, so $\mathrm{pr}(H/A)$ and $\mathrm{pr}(H/{\sim}A)$ are nearly the same. This is quite unintuitive; *A* seems just as good a sign of deer as it was before *D* became highly probable. Moreover, once the agent finds the antler and *A* becomes very probable, even $S(D,H)$ becomes very low. Here, we have an even more counterintuitive case: the agent has quite reasonably come to have high confidence in *H*, based on learning *A* and *D*. But according to measure *S*, neither *A* nor *D* now gives significant support to *H* (though, oddly, their disjunction does)![22]

This seems, at first blush, to be much like the original problem of probable hypotheses. But it is certainly not precisely the same problem; we have seen that *S* does not automatically preclude strong confirmation of highly probable hypotheses. What, intuitively, is going wrong?

---

[21] In this situation, the probabilities for *H* conditional on ~*D* and on ~*A* will be close to the unconditional probability of *H*—around .5—conforming with the intuition that the fact that there were not, for example, droppings *at location x* would have little effect on the probability of deer living in the wood. But the probabilities for *H* conditional on *A* and on *D* will be close to 1. Subtracting the former conditional probabilities from the latter thus leaves *S*-support at about .5 in each case.

[22] It is worth noting that this problem is different than that described above for the historical approach. There, after both pieces of evidence were discovered, their evidential weights depended on an irrelevant factor: the order of discovery The problem here is different, though no less real.

Again, I think that it will be helpful to look at the example halfway through the story. Let us examine the first counterintuitive result—that the confirmatory power of the antler is drained when discovery of the droppings raises the agent's probability for deer. At this point, the low value measure $S$ assigns to $A$ reflects the fact that, *given the agent's probabilities*, $A$'s truth or falsity is not highly relevant to $H$. But if $A$ is not highly relevant to $H$, given the agent's probabilities, why do we have the intuition that $A$ is highly confirmatory? It seems to me that, when we think about the question of $A$'s evidential bearing on $H$, we are impressed by the fact that, given the agent's general background beliefs about the relations between deer, antlers, and droppings, we can see that, *putting aside the agent's specific evidence* $D$, $A$ would be highly relevant to $H$. The intuitive idea—admittedly a vague one—is that the agent's degree of confidence in $A$ would make a big difference to his confidence in $H$ ceteris paribus, where the ceteris paribus condition would hold the agent's general background beliefs constant, but abstract away from the agent's beliefs that are being thought of as his specific evidence about $H$. The question seems to be something like 'How important would belief in $A$ be, if it were the agent's only evidence about $H$'?[illegible]

---

[illegible] Maher's account of confirmation involves various contextual factors, among which is a specification of a set of evidential propositions "whose relevance to $H$ is under discussion or otherwise salient in the context" (*op. cit.*, p. 167). But Maher's approach differs markedly from the present one in a number of ways. First, his account is purely qualitative (Maher is unpersuaded by arguments that the standard account of qualitative confirmation is adequate). Thus it makes no attempt to address the quantitative issues with which we have been concerned. Also, Maher's account makes use of other factors which go beyond an agent's probability distribution.

Maher's analysis relativizes confirmation to (contextually determined) "background evidence," which is taken to be so complete that evidence $E$ confirms $H$ relative to that background only if $E$ is positively relevant to $H$ for *every* probability function that could be rational for an agent whose total evidence consisted of that background. The assumption that the relevant background can be contextually determined in this way makes Maher's account objective, in the sense that it is not relativized to an agent's probabilities at all. But this assumption would be rejected by those who see confirmation as depending on past learning which is encoded in an agent's degrees of belief in a way that does not allow one to isolate out the evidence on which those degrees of belief were based. Maher's account also depends on an unanalyzed determination of whether some propositions are "based on an inference from" others. To my mind, this last aspect of the account especially raises troubling questions, since it seems to be presupposing a notion that is fairly close to (and may itself depend on) the notion being analyzed. But more detailed discussion of Maher's account must await another occasion.

To test this diagnosis, it would be natural to construct a situation which is similar to the deer case, but in which the probability of *H* is high based on background alone. In such a situation, it should seem more reasonable to say that *A* is not strong evidence for *H*. But it is a bit hard to do this, for the following reason: what counts as background in a given context will depend on the hypothesis under discussion. If we are discussing *H*, then what intuitively counts as evidence will naturally include whatever supports the agent's high degree of belief in *H*. But this makes it hard to think of a case where *H* is probable on the basis of "background" beliefs alone. The following is, however, an attempt in this direction:

Let *H* be 'There are at least five trees in the world', and *A* be 'There are at least four trees in Jocko's back yard'. Suppose first that a person has ordinary beliefs about trees, and thus a very high probability for *H*, but does not know how many trees are in Jocko's yard. In such a situation, I think that it would be unnatural to say that *A* would, for that agent, be strong evidence for *H*. That is because *H* is so probable anyway, given the agent's background beliefs, that the agent's lack of confidence in *A* makes little difference. (Here measure *S* gives the intuitively reasonable verdict, since $\mathrm{pr}(H/E)$ is essentially the same as $\mathrm{pr}(H/{\sim}E)$.) Were the agent to become highly confident of *A*, it would still seem unnatural to say that *A* was good evidence of *H* for her (and *S* would still agree with intuition). My suggestion is that our intuitions in this case differ from those in the deer case because *H* is so probable, given the relevant "background" beliefs, that we do not see confidence in *A* as making a significant difference to confidence in *H* ceteris paribus. Measure *S* gets this one right (and the deer example wrong) because it essentially treats everything other than *E* as "background."

In sum, then, it turns out that measure *S* does possess a significant advantage over the traditional measure: it avoids the old-evidence problem (and its first cousin, the probable-hypotheses problem). *S* cannot be seen as an accurate model of confirmation, however. The reason does not ultimately stem from the particular features with which Earman was concerned, but rather stems from the fact that *S* is insensitive to an aspect of the way we think about confirmation: the distinction between specific evidence and background assumptions. The inability to account for this feature, of course, is not peculiar to *S*. It will affect any measure that tries to analyze confirmation based purely on an agent's degrees of belief.[34]

## IX. CONCLUSION

Formal models of philosophically interesting concepts are tested largely by seeing how well they match intuitive judgments in particular cases. Persistent failures to find a formal model that matches our intuitive judgments may mean that we have yet to find the technical notion that corresponds to our concept. But they may also indicate that there is some feature of the concept itself which resists formal capture.

In some cases, this can simply be due to indeterminateness. Perhaps the controversy between difference- and ratio-based positive relevance models of quantitative confirmation reflects a natural indeterminateness in the basic notion of "how much" one thing supports another. This sort of resistance to formal capture should not, of course, be seen as undermining the philosophical interest of the formal models. If the models can show us two different ways that an indeterminate concept can be made precise, then they have taught us something.

The problems presented by old evidence (and by probable hypotheses) are not, however, of this sort. Both problems threaten to undermine the usefulness of the traditional model of quantitative confirmation (as well as its ratio-based variants). And old evidence does the same for confirmation models based on likelihood ratios. In this respect, the fact that measure $S$ avoids those problems constitutes an important advantage.

But even measure $S$, as we have seen, does not succeed in matching our intuitive quantitative confirmation judgments. And the reasons for the mismatches suggest that no purely probabilistic account could match those judgments. The question then arises whether the fact that *S* avoids even an important problem like the old-evidence problem can have any philosophical importance.

---

[34] The examples considered in the text make clear that we could not avoid this problem even if we enriched our representation of a probability function in a way that would allow us to recover a "pure" initial probability function and each bit of evidence that had combined with that initial function to produce the agent's present function (see Brian Skyrms, "Three Ways to Give a Probability Assignment a Memory," in Earman, ed., *Testing Scientific Theories*, pp. 157-61). For the background beliefs which underlie the probabilistic connections between evidence and hypothesis—for example, that deer shed antlers and drop droppings—are themselves based on evidence. Using the "pure" function would dissolve these connections. On the other hand, using a function that resulted from combining the initial function with all the evidence except the particular item in question would result in seeing no significant confirmation in cases like the one involving both antlers and droppings.

It seems to me that it can. Even models that fail to capture some important features of ordinary conceptions may be quite useful in helping us understand those conceptions, and the phenomena to which those conceptions answer; getting the cases right is not the sole point of a philosophical model. To the extent that a formal model helps one to isolate, identify, and understand those aspects of confirmation which resist formal capture, the model renders important philosophical service.

*S* is best seen not as a correct account of confirmation, but as a useful tool for understanding confirmation. It models one of the central aspects of confirmation—the one involving *E* making *H* more probable. Moreover, it does this while avoiding some problems that limit the usefulness of the traditional probabilistic measure (and the other popular measures we have considered). Its avoidance of these problems has helped us see, by examining the very cases where *S* gives unintuitive verdicts, other aspects of our conception. It has long been clear that the traditional probabilistic account of confirmation would fail to capture the epistemic asymmetry of our ordinary notion. It now seems that probabilistic accounts will also miss our ordinary notion's dependence on the distinction between background beliefs and specific evidence.

I would not claim that these are the only aspects of the intuitive conception that the measure fails to match. But we should not hastily take further mismatches as vitiating the philosophical interest of probabilistic models of quantitative confirmation. For in examining the cases in which our formal model gives unintuitive results, and putting our fingers on the reasons for the mismatches, we may yet reveal to ourselves more clearly the contours of confirmation.

DAVID CHRISTENSEN

University of Vermont

# NOTHING, SOMETHING, INFINITY *

In a memorable passage, Bertrand Russell[1] instructs us that:

> There is an argument, suggested by a passage in Plato's *Parmenides*, to the effect that, if there is such a number as 1, then 1 has being; but 1 is not identical with being, and therefore 1 and being are two, and therefore there is such a number as 2, and 2 together with 1 and being gives us a class of three terms, and so on. This argument is fallacious (*ibid*, p. 138).

In spite of its alleged fallaciousness, I shall try to justify a form of the *Parmenides* argument. Specifically, I am after a proof of the conditional:

> (Inf) If there exists at least one thing in the world, there exist in the world infinitely many things

Read contrapositively, (Inf) asserts that, if the world does not have infinitely many things in it, there is nothing in it. We are left with a stark choice: there cannot be in the world only a mere something; it is either infinity or nothing.

## I. THE ONE-REALM PRINCIPLE

The formulation of (Inf) speaks of 'the world' and 'things' in it. This calls for some discussion of what I intend by these phrases.

Traditional metaphysical analyses distinguish, among beings, between *reals* and *ideals*. Reals (for example, orcas and diamonds) belong in the material world, often alluded to as the *first realm*. Ideals belong elsewhere, in "other realms." How many other such realms?

---

* In the last few years, I published a trio of interrelated papers. "Logic and the World," *Journal of Philosophical Logic*, XVIII (1989): 197-220, "The What and the How," this JOURNAL, LXXXVIII, 5 (May 1991): 225-44, and "The What and the How II," *Noûs*, XXX,4 (1996): 413-33. The trio investigated the kinds of things there are in the world and which, if any, the world must have The present paper is a sequel applying itself to the question of just how many things there must be in the world.

The paper has been delivered in the last four years at various institutions. I am grateful for these discussions. Individually, I owe thanks to Torin Alter, Philip Bricker, Tyler Burge, Edmund Gettier, John Heintz, Ignacio Jane, David Kaplan, Ali Kazmi, Calvin Normore, Seana Shiffrin, Sten Lindstrom, and Byeong-uk Yi. Special thanks for objections discussed in section IV are due to Guy Rohrbaugh, Andrew Hsu, Tony Martin, and Dominik Sklenar (on whose work on kinds, see below).

[1] *Introduction to Mathematical Philosophy* (London: Unwin, 1919). By 'term', Russell does not mean a linguistic item but an object (entity, thing).

0022-362X/99/9609/462–78

Some (for example, George Berkeley and David Hume) opt for an exclusively mental ("psychological") realm, often referred to nowadays as the *second realm*. Yet others (for example, Gottlob Frege) speak of an additional *third realm* of nonpsychological, reputedly objective, abstracta, such as thoughts (*Sinne*), functions, and "logical objects" (for example, numbers).

It is against the background of realm proliferation that we are to read our one *ur*-principle:

> (One Realm) There is only one realm of being, the real world.

Our *ur*-principle propounds a unification of all beings in one receptacle—the real universe—generated by the "Big Bang" a few billion years ago. Whatever is—be it material, mental, or abstract—is a real historical being, one that came to be and existed in this one cosmic realm.[2]

One realm, then. The proof of (Inf) is about it: if there is in the real universe at least one thing (for example, Nastassja Kinski), then there are in it infinitely many real things. But from Aristotle to Thomas Aquinas, from G. W. Leibniz to Immanuel Kant, and all the way to our logically sophisticated end of Millennium, it has been repeated that the concrete (sometimes: "the sensible") pluralities of the world are merely finite or, at any rate, not provably infinite. As long as our focus is pluralities of concreta (actresses, orca whales, or diamond stones), there is no proving their infinitude.

Let this much be granted: ($\mathrm{Inf}_C$), 'There are infinitely many concrete things', is unprovable. But what about our project, proving 'There are infinitely many real things' ($\mathrm{Inf}_R$)? It is presumed that the unprovability of ($\mathrm{Inf}_C$) transfers to that of ($\mathrm{Inf}_R$). What lies behind the exclusion of a category of nonconcrete ("abstract") worldly things?

*I.1. The worldly abstracta dilemma.* For our paradigm of concrete subjects, consider a mischievous Puget Sound orca, Godiva, and the resplendent diamond on Liz's ring, Shine. Each such concrete object is both (i) objective and (ii) worldly (*wirklich*). Call the conjunction of traits the *mundane conjunction*, with both meanings of 'mundane' in mind: the in-the-world and the run-of-the-mill character of the conjunction of traits. But mundane as the combination is for Shine and Godiva, it appears intrinsically impossible for the abstracta manifested by this pair.

---

[2] A more canonical gloss of the adjective 'worldly' ('real') shows up later in the paper (see 'the reality quintet'). At this point, I follow this 'do-not-define-it-use-it' procedure because I find the appeal of the notion—worldly, real—independent of my own specific annotations. For example, Frege's gloss of 'real' ('*wirklich*') suits us very well. "The world of the real is a world in which this acts on that, changes it and again experiences reactions itself and is changed by them. All this is a process in time. We will hardly recognize what is timeless and unchangeable as real"—"The Thought," in P. F. Strawson, ed., *Philosophical Logic* (New York: Oxford, 1967) p. 37.

With these abstracta, it is either/or. Make it (the number Two, the species Orca, the substance Diamond, the colors Black and Blue) objective, as Platonists have, and it is banished away to another, otherworldly and acosmic, third realm. Alternatively, we may follow the nominalist and (like, for example, John Locke) make each such abstractum into an individual mental abstrac*tion*. Less "individualistically," we may view such abstracta as essentially dependent on our linguistic use of the classificatory predicates '...are two' and '...is an orca' and '...is blue'. Either way, we have brought the abstractum back into the world—the psycholinguistic second realm—but at the cost of losing its objectivity.

This then is our problem: the abstractum (say, the species Orca) cannot co-exist in the first cosmic realm with the concreta (Godiva, Gieronimo, and the like) that realize it, make it real. The problem is metaphysical and is not to be confused with Paul Benacerraf's dilemma, which is essentially epistemic. Our problem does not concern sensory access and forms of knowledge; it arises for an omniscient being like God: even He cannot make abstract objects bear the mundane conjunction, be both objective and in the cosmos. Even for Him it is either/or, what I shall call the *worldly abstracta dilemma*.

*I.2. The worldly infinity dilemma*. The worldly abstracta dilemma is our root problem. It leads to a derivative quandary: no plurality of things may be both of worldly constituents and yet contain infinitely many items. It is again either/or. The choice we are forced to make may be impressed on us by considering two classical arguments for (Inf), one due to Frege, the other to Richard Dedekind.

Frege's argument first. He begins with an ideal, off-world item, the predicative concept (Frege speaking here) 'not identical to itself', $F_0$. We now assign $F_0$ a yet higher (second) level predicative concept, a number concept, ZERO. Still on logical grounds, we now "take" the extension of ZERO. This is by Frege's lights a pure "logical object," the number 0. We existentially generalize: there exists at least one object and 0 is it. Frege now ascends back to the realm of predicates and considers this time the concept $F_1$, 'be identical to 0'. Correlated with it is a number concept (namely, ONE) the number of things falling under $F_1$. The extension of ONE is a new object, the number 1. By now, Frege can prove the existence of two logical objects, 0 and 1. And so it goes: the process yields infinitely many items when Frege proves by induction on $n$ that the extension of 'number of items smaller or equal to $n$' is always $n+1$.

What has Frege proved? The infinitude of what he calls *logical objects*, predicative concept derivatives, items that have it built into their very identity that they are extensions of concepts. Therein lies Frege's version of our dilemma: insist on *wirklich* subjects, and no proof of (Inf) is forthcoming; shift to off-world predicative abstracta, and a proof becomes available.

A similar message is imparted by the main other classical proof I am familiar with, Dedekind's. In section 66 of his *Was Sind und Was Sollen die Zahlen?*[3] Dedekind tries to prove that there exists at least one infinite "system" of objects. On his way, he argues that there is an infinite plurality of *objects*.

Dedekind locates the infinite plurality in his *Gedankenwelt*. His *ur*-object is his Ego. He then goes on to apply to this realm of thought a one-one map $f$ that is not onto. The informal gloss of $f$ is 'can be the object of my thought'. We are told without further ado that $f$ is one-one. And obviously it is not onto: Ego is not got as a result of the map. This much makes the realm of these *Gedanken*-objects Dedekind-infinite, hence infinite simpliciter. But there remains the question: What precisely is proved by Dedekind?

Russell's commentary says it all:

> We are then to suppose that starting say with Socrates, there is the idea of Socrates, and so on *ad inf.* Now it is plain that this is not the case in the sense that all these ideas have actual empirical existence in people's minds. Beyond the third or fourth stage they become mythical. If the argument is to be upheld, the "ideas" intended must be Platonic ideas laid up in heaven, for certainly they are not on earth (*op. cit.*, p. 139).

The worldly infinity dilemma, all over again.

## II. DEFUSING THE WORLDLY ABSTRACTA DILEMMA[4]

Consider again the map used in Dedekind's proof. Instead of his original 'can be the object of my thought', substitute 'set of'. This operation is one-one: if $x \neq y$, then $\{x\} \neq \{y\}$; and it is not onto; whatever it exactly is, 'meine eigenes Ich' is surely not a set. We thus have on hand a (Dedekind)-infinite plurality of real items. Was the above allusion to dilemmas much ado about nothing?

I shall not rest my case on the category of sets. Granted, sets, as characterized by Dedekind, Georg Cantor, and eventually Ernest

---

[3] Braunschweig Viewey, 1960, eighth edition.

[4] I am grateful in this section to Andrew Hsu and Guy Rohrbaugh.

Zermelo, are not mere predicative projections, like Frege's extensions or Russell's propositional functions. Nonetheless, sets are not what I here target by 'real things'. It is thus time to set out in detail what I intend by real (worldly, cosmological, *wirklich*) items, why sets will not do and which abstracta will.

Five marks of the real suggest themselves, what I shall call henceforth the *reality quintet*:

(1) No prefabricated individuation
(2) No reduced identity
(3) Historicity I: constitutional change
(4) Historicity II: predicative change
(5) Historicity III: counterfactual developments

Consider first the principle of no prefabricated individuation criteria. With real items (for example, Godiva), it is the actual existence of the entity, its coming into history and development in it that fixes whatever is distinctive about it. No set of antecedent conditions—prior to the historical item—may define or dictate what it is to be this specific orca whale Godiva.

A very different reading of individuation principles applies to sets and Goodmanian sums. Here, we do have axioms of extensionality that *regulate* in advance what it is to be this specific construction (the set *X* or the sum *Y*) in terms of members or parts. The fact that such a principle is formulable for sets and *G*-sums (as for other "constructions" like aggregates, tropes, and the like) is, for me, an indication that we do not have on hand a genuinely historical subject, produced by the autonomous forces of cosmic evolution.

Let us now turn to a third category of items: for example, the species Orca and the Manchester United soccer team. Do we get the profile of Shine and Godiva, or that of sets and *G*-sums? The species and the soccer team are historically real phenomena: it is only their actual emergence in history that fixed the distinct identity of the item that emerged. We date the emergence, namely, the *origination* of the species and the *founding* of the team: the species did not exist five billion years ago and the team did not exist in René Descartes's time. No prefabricated set of criteria could legislate into existence this species or this soccer team.

The second mark—no reduced identity—is related. A set (Goodmanian sum, aggregate of parts, trope) just is (is identical to) an ensemble of "atoms" making it up. As such, the identity of the "higher" entity is reducible to a constructional operation of "composition" (set of, aggregate of, and so on) on more basic ingredients. In contrast, Godiva and Shine are not so reducible to "constructions" out of ingredients. We find the species and the soccer team sharing this profile. Even though both the species and the soccer team have a multiplicity of material "members," neither is identical to a logical construction out of the membership.

This logical point regarding identity is manifested over the temporal dimension, when we consider the manner in which the pertinent entities endure in history. The set and $G$-sum would not exist, if it were not for their "ingredients" (for example, {*Godiva, Shine*} did not exist before planet earth did). But once in history, the continuing existence of the assembled entity is tied intrinsically to this specific constitution; there is no way for this set or $G$-sum to persist except through the persistence of this specific constitution.

Not so for Godiva and Shine. Their survival through history depends on constitutional change: the higher entity goes on existing, as its constituting matter changes, only because it is realized in a different material basis. Again, the species and the soccer team follow this mould: the evolution in time of this single species and soccer team *presupposes* changes in membership; it is only because the higher entity can sustain itself by having other members that *it* endures.

Constitutional change leads us to another form of change: predicative change. Real subjects like Godiva and Shine constantly change the properties they bear. Such is also the profile of the species and the soccer team: each ages, evolves, scatters, becomes endangered—all properties borne by the higher subject and not reducible to properties borne individually by its present members. Not so with constructions. Doubtless, the set-theoretic pair may change: it had a member (Godiva) that was hungry and now it has a member that is bellyful. But such predicative changes are essentially by proxy, analyzable via changes in the intrinsic ingredients.

The availability of an *actual* history leads us to a final characteristic mark of matters real—the potential for alternative developments of that history. Such counterfactual developments are surely envisageable for Godiva and Shine and, in turn, for the species and the team. No such counterfactual developments are open to abstracta of the constructional kind. Just as they lack a genuine *actual* history, they are not subject to alternative evolutions of that history.

Our reality quintet sets apart "blueprint"—constructed—abstracta from those which are historically real. Two observations are suggested by this separation. First, we have seen examples (the species Orca and the Manchester United soccer team) of genuine worldly abstracta, bearers of the mundane conjunction; each is both objective and cosmic. Our root dilemma thus turns to be a false one. Second, regarding the worldly infinity dilemma, we know what to look for: items which are in the world like soccer teams and species but of which there are infinitely many.

## III. INFINITY IN THE WORLD

What is proved below is a form of the *Parmenides* argument.[5] We are given the existence of one (purportedly the first) concrete item: the orca Godiva.[6] By its very primal existence, Godiva generates a variety of real kinds: the color-kind $^k$Black$^k$, the shape-kind $^k$Oval$^k$, the biological species $^k$Orca$^k$, the kind of bodily-organ $^k$Brain$^k$. One further kind generated by Godiva is the arithmetical kind $^k$One$^k$. She is most definitely a black thing, an orca-thing, and so on, but she is also one thing; at that, the *first* among all ones.

So now we have on hand an ordered two-some: Godiva and an item she generated, $^k$One$^k$. Both Godiva and $^k$One$^k$ exist—two real things. Thus, the first real instance of $^k$Two$^k$ exists; in turn, $^k$Two$^k$ is generated. We now have on hand three ordered, real things: Godiva, $^k$One$^k$, and $^k$Two$^k$. The kind $^k$Three$^k$ is generated.

The foregoing generation of the finite ordinals produces a plurality of the following form: Godiva, $^k$One$^k$, $^k$Two$^k$,... The plurality has $\omega$-many members. This generates the first infinite ordinal, the kind $^k\omega^k$. Once it exists, we have on hand the plurality: $^k$One$^k$, $^k$Two$^k$,..., $^k\omega^k$. Here lies the generative basis for the ordinal $\omega+1$. Past the generation of ˹One˺, we use two further generative principles—the very two used by Cantor (before he gave up on ordinals as primitive and reduced them to sets): the addition of one item to a given ordinal plurality and the "taking of a limit" of an unending plurality of ordinals, as in the case of ˹ω˺.

---

[5] The idea of "proving" any variant of (Inf) may seem passé to many a reader; we should rather state it as an "axiom" of infinity. Let me clarify what is here meant by "proof."

First, I am not out to prove that there exists at least one infinite system (set, proper class), and so on. Second, neither Dedekind nor the present account aims to prove (Inf) inside a formal (be it generalized first- or second-order) logic deductive system. It is a familiar fact that such formal logic systems (where (Inf) is expressible) afford models with finite domains.

What is offered here is an informal proof that (Inf)—at that (Inf$_k$)—is actually true. This much is not more ambitious than what is pursued by standard set-theoretic axiomatizations. Consider such an axiomatization, say, standard ZF. The author of the axiomatization (textbook) explains to the student why the axioms are *true*. This much applies also to the axioms of pairing, union, and so on. Such an explanation must also be given for (I)—there exists an infinite set. After all, set theories explicitly denying (I) are available. So, do we or do we not give a "proof" of (I)?

The answer is that most careful authors back up their assertion of (I) with the display of an infinite set—usually the one representing the natural numbers in the Zermelo or von Neumann systems. This is the textbook de facto proof. Indeed, such proofs seek a more ambitious result than ours. For it is argued that the aforementioned infinite plurality of finite sets forms in turn a single infinite set. Our argument does not take the last step (though I do not believe there is anything wrong with it).

[6] In what follows, I use the notation $^k$Orca$^k$ to make clear I speak of the entity, the zoological kind of Orcas.

*III.1. Annotations to the argument. (i) Generation versus instantiation.* We should separate the item(s) that bring a kind into being from item(s) that, once the kind exists, count as its instances. Suppose Adam is the first being in the world. His existence generates that of ˹One˺. Later, Eve shows up. She counts as an instance of ˹One˺, not as its generator. And so it goes. When Adam generates ˹One˺, he and ˹One˺, the two things that they are, generate ˹Two˺. Later, Eve shows up. Adam and Eve, the two-some plurality they make, provide an instance of ˹Two˺. Still, they are not the generators of ˹Two˺, a role reserved for Adam and ˹One˺.

On the present account, any *n*-some plurality, whether of numbers, cabbages, or kings, may serve as an instantiation of ˹*n*˺. But only the preceding sequence of *ordinal numbers*, up to ˹*n*˺, can generate ˹*n*˺. In all, it is suggested both that (i) only a non-number, a prenumber concrete object, can generate ˹One˺ and (ii) from then on, only numbers can generate *their* successors.[7]

*(ii) Ordinals versus cardinals.* The procedure described above generates first the (finite) ordinals. The ordinal is the prior *existent*. The equivalence class that it is, cardinal Card(˹Two˺) rests on the subsequent *truth* of the predication: 'plurality having as many items as the one that generated this existent, the ordinal (˹Two˺)'.

In this respect, Card(˹Two˺) is rather like the biological kind-predicate 'is an orca' which I read as 'is an instance of the kind ˹Orca˺'. In turn, the objectual ordinal kind ˹Two˺ is the analog of the objectual kind ˹Orca˺. In both cases, the kind-as-object precedes the predicate 'being a member of that kind'.[8]

---

[7] On both (i) and (ii), see more in the annotations that follow.

[8] In the context of infinite numbers, the difference between generation and instantiation, ordinals and cardinals, is quite clear. The ordinal ˹ω˺ is generated first, then ˹ω+1˺. By quite a different existence principle, ˹ω˺ gives rise to the cardinal number $\aleph_0$. Now, the instances of ˹ω˺ and ˹ω+1˺ are quite different: all pluralities of the order type ω are instances of the former; one more item in the plurality or an order change will be required to make a plurality that instantiates the second ordinal kind. But now when it comes to instances of the first infinite cardinal, all such instances, regardless of their order type, would do as instances.

The ordinal/cardinal distinction shows up clearly for infinite numbers; it also shows up with Zero. On the present account, Zero is not an ordinal. What "defines" Zero into existence is predication: it is the number of things falling under empty predicates. With, for example, Card(˹Two˺), we test for one-one correspon-

*(iii) The reality quintet. (iii.1) Identity criteria.* First on our list of marks was the subject's affording *no prefabricated individuation.* Classical conceptions of number supply such individuation criteria in, for example, the Frege-Russell vein: the number of *F*s and the number of *G*s are one and the same if and only if the classes (concepts) of *F*s and *G*s are in one-one correspondence (lately known, under the influence of the late George Boolos, as "Hume's principle" and universally viewed as "logical," or at the very least "analytic" and "defining").

On the present account, it may turn out of such a (biological, chemical, arithmetic, and the like) kind that some fundamental feature of its generation process is uniquely true of it; thus, different persons (rivers, species, chemical elements, numbers), different generative histories. But these are ex post facto *historical facts,* not prefabricated regulative conditions. It may thus well turn out that different ordinal numbers have different generative histories (different numbers, different ancestors) but this much is simply a description of how each came to be, not a stipulated convention.

*(iii.2) No reduced identity.* The reduction would amount to the number's being just a certain ensemble of more basic ingredients (parts, sets, concepts). Our number-kinds are not reducible to their members. The existence of members is critical for the generation and ongoing existence of the kind but the kind itself is never exhausted by this or that membership.

*(iii.3) Historicity I: constitution change.* Consider the difference between our kind 'Two' and, for example, von Neumann's set 'Two'= {*Godiva, One*} (*s*-quotes are to remind us the object is a set). First, regarding identity: von Neumann's 'Two' is a *specific* two-membered set. Benacerraf's other problem—"What numbers could not be"—strikes: Why is the number Two identified with this set and not any other two-membered set? In contrast, our kind 'Two' is at the right level of generality: it comprehends all two-item pluralities as instances (the instances of the kind are two-some pluralities (for exam-

---

dence with the *ur*-two-some plurality that generated the ordinal 'Two'. In the case of Zero, we cannot so test, for there *exists* no such *ur*-exemplar and no generated ordinal. All there is is the one-one correspondence between all empty predicates. We may feel here, with Frege, that Zero is an adjective that has grown capitals (for example, 'there are no(zero) unicorns').

The foregoing relates to a referee's question whether we could use our real-number-kinds (for example, 'Two') to number third-realm items (for example, two forms in Plato's heaven or two Fregean extensions) The answer tracks back to my above (one realm) principle: all beings exist in this one real cosmological realm. What is not—is not *wirklich*—cannot be numbered. Only reals can enter real relations—*be numbered by*—with a real like the number 'Two'. On which, see more in annotation (*iii.4*).

ple, me and you), not the subsequent unity (the two-membered set {*me, you*})).

Second, regarding existence. von Neumann's ‘Two’ is perilously dependent on the existence of Godiva: no Godiva, no ‘Two’. In contrast, the kind ‘Two’, like ‘Orca’, goes on existing while its members change. The number-kinds (again, like biological ones) were *generated* by a specific *ur*-item—the historically first concretum. But to go on existing, all they need is to be *instantiated* by some—any old—instances.[*]

*(iii.4) Historicity II: predicative change.* ‘Orca’ and ‘Diamond’ change through history. But do numbers ever change? At first blush, the difference between the cases seems striking. But let us have a closer look. Granted, 9 is fixedly odd and greater than 7. But this fixidity, just as with the temporally stable features of ‘Orca’, concerns features fixed by (and at the time of) the very generation of the item. On the other hand, 9 also numbers the coins in my pocket. This is a feature gained by 9 after its generation and thus subject to change. When we say the next day ‘9 used to number the coins in my pocket but it no longer does’, we make a claim of *de re change*: some specific *res* was one way, now it is not. The question is: What *res* is that?

The natural answer seems to me: the number of the coins. Of course, at first blush, we think the change is, to speak loosely, "in the coins," not in the number. But what does it mean the change is "in the coins"? It could not be a de re change in the nine membered *set* {*c*1, *c*2, ..., *c*9}. This set cannot survive the addition or subtraction of a coin. Nor is it a change in the aggregate or the Goodmanian sum. I submit that the change in the coins induces a change in the number but it is not identical to it—the change in the number may have taken place in other ways (instead of adding *c*10, I may have added

---

[*] What if we added to the reality quintet a sixth condition: all reals must at one time or another cease to be? Would this not make the existence of an infinity of our kind of numbers ephemeral?

My answer is twofold. First, quite apart of numbers (and other real abstracta), I would not add the purported sixth mark of being real. Generation—as a mark of the real—is not "symmetric" with corruption. Many generated abstracta (including individuals, not just kinds) may go on existing by replenishing the materials that realize them. This gives such real subjects—as opposed to sets of such material ingredients—a long lease on life.

Second, regarding numbers. Once there was one concrete thing, there were numbers. And if—*per impossibile*—there were nothing concrete, there would be nothing at all. As will become clear in the next annotation, I believe that once there existed something concrete, there always (and of necessity) continued to be some concrete item In time, the infinite generating *sequence*, with the specific Godiva at its basis, will be gone, because Godiva will corrupt. But the numbers (the second place in this sequence, ‘One’, the third, ‘Two’, ) are forever.

*c*11). Likewise, there may have been, in the concrete coins, the following change: the addition of one blue coin *c*10 (the original nine are red). But it would be a mistake to identify the change thus induced in the *number* of coins with the change in the *color*. Although actually realized via the addition of *c*10, the change in number and color might have occurred separately.

One may question the foregoing emphasis on the intuition of de re change. Thus runs what I call *the no-change objection*:

> Consider "The president used to be George Bush and now it is Bill Clinton." Rather than a change in the state of a single underlying entity, we have *replacement* of one entity by another: Bush was the president and another individual, Clinton, is now the president. In a similar vein, we may suggest for the above claims that no number, color, and so on has really changed. We have instead replacement: one color, green, was the leaves' in the spring, now it is another, yellow; one number, 9, was replaced by another 10, as the number of coins in my pocket.[19]

In response to the no-change objection, I shall say this. First, we should separate in this discussion syntactic and metaphysical issues. Grammatically, we have on hand a uniformity between 'Bert's girlfriend (the U. S. president) is changing', 'The color(number) of the coins is changing'. Metaphysically, going by the present paper's outlook, species, temperatures, colors, and numbers are real kinds, coming in and out of cosmic history independently of our verbal activities. In contrast, the "roles" of being the American president and Bert's girlfriend are induced by *predications* of ordinary individuals. I shall attend here to real kinds and leave to the aforementioned linguistic companion piece the proper analysis of 'Bert's girlfriend (the president) is changing'.

Second, let us ask now: When I speak of "the number (color) of coins is changing" as involving de re change, in *which* res is the change to be located?

---

[19] The *no-change objection* has been made to me by a variety of commentators in the University of California/Los Angeles Language Workshop dedicated to the Partee-Montague puzzle regarding the sentence 'The temperature of LA is ninety and it is rising'. It was also made by François Recanati.

I view Montague's original analysis of 'The temperature is rising' as in the no-change objection vein. He lets the prime semantic value of 'the temperature' be a function from times to numbers. He then speaks of that function as "changing" ("rising"). In truth, the function does not change (rise) at all. It is as fixed as any thing could be, the set of ordered time-number pairs that it is. What is actually encoded by the function is *replacement*: the replacement of one extension (90 degrees) by another (91 degrees) as the extension of 'the temperature'. I discuss in detail Montague's analysis, the no-change objection, and the syntax-semantics of such constructions in my "The Subject-intransitive Verb Class" (in press).

The question cannot be answered without understanding the *pre*change kind of fact reported in the likes of 'The species of Godiva is orca' or 'The color of leaves is green'. Some (for example, Richard Montague) regard the claims as having the form of identity statements. With this, I disagree. We have identity claims in the likes of 'Furze is Gorse' (species identity), the bilingual 'Green is Verde' (color identity), '34=XXXIV' (number identity).

In contrast to an identity, 'The species of Godiva is Orca' has the two-sorted relational form:

$$R(c,A)$$

where $c$ indicates the concrete relatum, $A$ the abstract relatum. Idem for 'The number of coins is 9', whose form I view as the relational: *Has-number* (coins in my pocket, 9).

I read much importance into the genitive form of our constructions, namely, the abstractum mentioned is *of*-the-concretum (a): its color, temperature, species, or, in the plural, their number. Generalizing, I would like to call such assignments of a concretum-to-a-kind ***kind-membership reports.***

Now, when we report change in such membership relations, we report change in a *relation.* The primary change is in the concrete relatum. But its (their) failure to stay related—on a given dimension (color, temperature, number, and so on)—to the kind (red, 90 degrees Fahrenheit, 9, and so on) induces a change in this second relatum—the kind. Here, the change regards what the color, temperature, and number are *of.*[11]

*(iii.5) Historicity III: counterfactual developments.* Nine actually numbers the coins in my pocket but it might have failed to number them, just as Nastassja Kinski was actually the leading actress of *Tess* but she might have failed to be if she resisted Roman Polanski's pressures. What makes these 'might have' claims true?

'$x$ might have failed to $F$' is true at a moment $t$ in the actual history of the world if and only if there is in the actual history of $x$ an earlier moment $t^*$ at which $x$ could have gone on to be not $F$. There are such moments—pre-1976 for Nastassja and prior to my filling my pocket with these nine coins for the parking meter. For concrete subjects like her, the necessity of bearing a trait would amount to there being no way *her* history could have gone otherwise. And in the same vein, this is our grounding of arithmetic necessities: although 9 might not have numbered the coins in my pocket, it could not have failed—at any point in its history—to be odd or greater than 7. Not only do we unify Nastassja and Nine in the same cosmic realm of existents; we also apply a unitary notion of *de re necessity*—cosmic necessity—to both Nastassja's being human and 9's being odd.

---

[11] It has been sometimes objected to me that this is a mere "Cambridge change" in the species, color, and so on. Not on the present view of kinds. We saw earlier that, for example, species identity and existence depend on (i) *what* animals the species was generated by, and on (ii) *there being* such animals to generate it. In contrast, when I Cambridge-change because you had a hamburger for dinner, nothing relating to my very existence and identity depends on your feeding behavior.

*(iv) The necessary existence of infinitely many things.* The idea of cosmological necessity assumes through and through that there is only one world—the real world. For a trait *F* to be necessary of the real world, we look for all the possible developments of the world past the moment of its actual origination. Now, looking at Nastassja, we do not consider among her alternative developments, a consistent story about *some* individual looking like Nastassja and having the same genetic endowment as she does but living, to begin with, in Descartes's time. In the same vein, we do not consider among the alternative developments of the history of this cosmos, consistent stories about *some* universe that would look (by now) like ours. A story that makes the world have a radically different history to begin with (for example, a Big Bang 500 billion years ago, a substantially different rate of expansion, or, following Aristotelians, no beginning moment at all) is just that—a story ("theory").

It is important to separate the conception of cosmic necessity from the more familiar "parallel worlds" conception of necessity. A sense of the difference in the case of numbers may be gleaned from the analogous and much rehearsed example of God.

On the "parallel worlds" conception, God's necessary existence is secured by his antecedent acosmic essence (definition). A plurality of possible worlds is given of which this historically real world is just one (one among many of a kind). It now turns out that in any arbitrary world *w*, something bears the divine essence at *w* and that something exists at *w*; furthermore, what bears it in some world *w*, bears it in every other world (shades of the ontological argument).

In contrast, the cosmological necessity of God's existence does not proceed from essence to transworld existence. We rather seek to establish why, among actual existents, His is special: whereas other real existents may be subtracted from the history of the real world, He stands in a special position in the actual cosmic history and may not be subtracted. His necessity lies in His insubtractibility from the history of this one and only world of ours.

So it is for God and so it is for the numbers. On the "parallel worlds" conception, our Parmenidean argument secures it that in any arbitrary world *w*, if there is one thing, there are in *w* infinitely many things. But, alas, I do not know how to argue for "parallel worlds" (given disjointly and point-like) that, to begin with, each must have something in it.

On the cosmological conception, we start with the actual fact that the world itself *is*. This world is not a "point" or an index of evaluation or a set of propositions; it is simply this real evolving universe. So, something is. But is its existence necessary? There is a point in the history of the world in which it could have gone on to develop without Nastassja Kinski, the Nanga Parbat, or planet earth. But is there a point *in* the history of the world at which it could have gone to be such that it did not exist in the first place? With the necessity of some real thing thus secured, our Parmenidean argument carries us all the way to the necessity of infinitely many reals.

*(v) The generation of ‘One’*. Some readers will want to grant that the sheer emergence of Godiva generates certain kinds (for example, ‘Orca’) but not the purported kind ‘One’, simply because this last is not a genuine kind. There is no genuine "common nature" or "unifying essence" behind all ones.

The present account of real kinds (for example, of biology and chemistry) is this. A process in history produces items of the pertinent kind and it is thus—and only thus—that the species and the substance come into history. Such is the worldly picture of ‘Orca’ and ‘Diamond’ and such is the intended account of ‘One’.

First, a concrete existent brings into the world this other, more abstract, kind of existent, the ordinal ‘One’. Eventually, all numbers are so generated, not "from above," by a third or second realm, defining predication but "from below," by in-history, preceding real existents. There is an intrinsic difference between (i) the first number, ‘One’, which can only be generated by a non-number, and (ii) all remaining numbers which can be generated only by other numbers, each by its predecessor. We should note that in (i)-(ii), we have a nice metaphysical reflection of the familiar (Peano-Dedekind) axioms: the *ur*-number (for Dedekind and for us) 1 is intrinsically *not* a successor; all others are intrinsically successors.

In the present framework, predecessor numbers actually precede the next number in the historical order of existence. Thus, the familiar fact that 16 is "after" (succeeds) 12 is not a truth got "from above," as if after 16 and 12 have been separately defined by predication, it *also* turned out that a plurality with the cardinal number 16 has "more items" than one with 12 items. When it comes to our generated ordinals, we have here a "from below" fact about the historical order of generation: 12 is among the ancestors of 16. This is meant not as a "metaphor" or "logical analog" of real ancestry relations. Not at all. Without the existence of 12 in history, 16 would not emerge.

One final repartee and a forceful one at that—the *separate beginnings objection*. Take to heart the foregoing claim that 'One' is as generable as ordinary natural kinds, and it is from the frying pan into the fire:

> Consider the spatiotemporally separate Dobbin and Bertrand, the first horse-like (in looks and DNA) animals. By your (JA's) own lights, the two generate two distinct biological kinds. But you say that number-kinds are like biological kinds. So, either you predict that (i) the world must have at most one primal concrete item or else (ii) if there were two (or more) simultaneous primal objects (for example, Electra (E) and Fidelia (F)), the two would breed two distinct kinds 'One'. This is an odd result.[12]

The source of the separate beginnings objection lies in misinterpreting the analogy between arithmetic and biological kinds. I would like to explain in what way they are similar and in what way they are not.

Biological kinds (for example, 'Horses') make prominent two separate features:

> (Generability) For a kind to exist, it has to have concrete generators.
> (Membership connectedness). To be a (late) member of kind *K* (for example, a contemporary horse), *x* has to come from *K*s (for example, from horses).

The point to notice is that (generability) and (membership connectedness) are logically independent. I definitely asserted above (generability) for any real kind whatsoever. But is (membership connectedness) also true of kinds in general or is it something peculiar (even if essential to) biological species?

One might be led to think that I submit (membership connectedness) for all kinds simply by confusing it with (generability). But even without such a confusion, one may reason as follows regarding the separate question of kind-membership.

---

[12] The objection that follows has been made to me early on by Martin, then by Normore, Shiffrin, and Dominik Sklenar at UCLA and Jane in Barcelona. The discussion of kinds that follows owes much to conversations with Sklenar about his 1996 UCLA Ph.D. "Being of a Kind"; a published version of his interesting ideas is forthcoming.

I should like it noted that the objection has the form of a disjunction: either the world must have at most one primal item or else if there were more than one such, then.... I believe the first disjunct is true. The remarks in the previous annotation about the necessary existence of the world itself give a hint of the argument I would submit. But in what follows, I shall assume with my opponent that there could be more than one primal item.

The critical question to consider is this: What unifies two items *x*,*y* in a given kind *K*? Either the unification is "from below" or "from above." If it is "from below"—due to a cosmological fact—the model must come from the unification operative in biological species, namely, via historical connectedness between members. If not, the only alternative unifier is an idea (*Sinn*, form) in another "realm" (for example, Locke's second, Frege's third, or Plato's "heaven"). This either/or I shall call the ***unification dilemma***.

Driven by the dilemma, our objector may well reason thusly: the present approach would not allow unification by a second-or-third realm template. Therefore, the one intended must be unification as in the biological case—by historical connectedness. By hypothesis, there is no connectedness between Electra and Fidelia. Thus, we are led to two number-kinds 'One'.

My answer is that the unification dilemma is a false dilemma. There are other "from below" ways (aside of historical connectedness) of unifying things in the cosmos. A pertinent example is provided by chemical elements, to which I now turn.

What makes any old atom *x* a member of the kind 'Germanium'? Let us assume that the emergence of this chemical element in the cosmos (so many (many) seconds after the Big Bang) is the result of two simultaneous but causally separate supernovas. Each supernova produces from electrons and protons by a similar process an atom with the structure of atomic number 72. The one Germanium atom call $g^*$, the other atom $g^{**}$. In my view, both are of the kind 'Germanium'. Yet surely, they are not historically connected.

What unifies $g^*$ and $g^{**}$ in the kind is neither (i) historical connection nor (ii) satisfaction of an apriori third-(second-) realm template. What makes an atom a Germanium atom is the type of cosmic process by which it came to be: electrons/protons at the right temperatures (and so on) coming together to form a unity.

Let me draw from the foregoing a general conclusion. What matters for membership at the general level of all kinds is this:

> (*M*) what makes *x* and *y* co-specific is that the processes by which they each came to be are fundamentally similar (in the pertinent respect: chemical, biological, and so on) and essential to *x*(*y*)'s very coming into being.

(*M*) is true of the historically bound tigers *x*,*y*; the historically disconnected germanium atoms *g** and *g***; and lastly, the separate simultaneous primals Electra and Fidelia.

It is essential to each and every object's (including Electra and Fidelia) process of coming into being that it would come to have unity, that it would become one thing. The coming together of the unity might involve the bonding of electrons and protons, DNA molecules, stars in a galaxy, and other more complicated modes of formation, depending on the specific kind of item concerned. But over and above specificities due to varying kinds, each object's very coming into being has essentially this much in it: the coming into being of one thing. We might well say: no entity without *numerical* identity. Physically, Electra and Fidelia may or may not be of the same kind. But this much is given about the two as about any other emergent being: Electra and Fidelia are of a common *arithmetic* kind, the one and only ⌜One⌝, because in both we have the fundamentally similar process of a coming into being of a distinct new entity.

## IV EPILOGUE

I have argued that there can be no existence of a mere something, be it a single initial something or a plurality of such primals. Each such primal being (as any other object) cannot exist without being one thing, a number that itself could not come into being without having a generating instance, some one concrete thing. So: no number one without a concrete generator; but also, no such concrete generator without the number one. Very well then: coming into the world with each such primal entity is the number one; and with that two-some, the primal entity and one, comes into the world the number two; and so on. We can sum it all thusly: there is no real infinity without real being and no real being without real infinity.

JOSEPH ALMOG

University of California/Los Angeles

# PUTNAM'S MODEL-THEORETIC ARGUMENT RECONSTRUCTED *

Metaphysical realism has been the main subject of the critical part of Hilary Putnam's work for over twenty years now. At the heart of this doctrine, as he presents it, is a thesis I shall call *methodological fallibilism* (MF): even an epistemically ideal theory may fall short of truth, where the notion of truth involved is that of *correspondence truth* (CT).[1] Note that MF is not merely about epistemically ideal theories. The claim is that *no* empirical theory, no matter how good our evidence for it (that is, not even if it should be epistemically ideal), is guaranteed to be true. A further characteristic of metaphysical realism is a view of semantics sometimes called *semantic naturalism* (SN). Its credo is that "Semantics is an empirical science like any other."[2] One consequence of this view is that theories of reference (theories of the correspondence that, according to CT, has to obtain between reality and a sentence for that sentence to be true) are empirical theories, and are thus to be accepted or rejected on the basis of exactly the same criteria that govern theory choice in what we more traditionally think of as the natural sciences.

Over the years, Putnam has advanced several arguments against metaphysical realism.[3] Of these, the so-called model-theoretic argu-

---

* I am grateful to Leon Horsten and Bas van Fraassen for helpful discussions on the subject matter of this paper

[1] As spelled out, for instance, in Michael Devitt, *Realism and Truth* (Cambridge: Blackwell, 1991), p. 29 Some passages in Putnam's writings suggest that he thinks MF to be a mere consequence of the realist's view on truth—see, for instance, his *Meaning and the Moral Sciences* (Boston: Routledge, 1978), p. 125. But from the fact that the metaphysical realist separates truth from any epistemic notion (like verification or rational acceptability), and in that sense conceives of truth as "radically non-epistemic" (*ibid.*), nothing follows about the (possible or necessary) truth or falsity of (ideal) empirical theories, at least not as long as we are not informed about the metaphysical realist's view on the relation between truth and evidence. That for the realist an ideally confirmed theory is not true by definition (as it is on some pragmatist conceptions of truth), and hence not necessarily true for that reason, does not mean that a realist could not, on other grounds, consistently come to the conclusion that an ideal empirical theory must be true. As it happens to be, methodological considerations—such as the underdetermination problem—suggest that this is not so, and thus motivate MF (cf. Bas van Fraassen, *The Scientific Image* (New York: Oxford, 1980); John Earman, "Underdetermination, Realism, and Reason," *Midwest Studies in Philosophy*, XVIII (1993): 19-38; and Leon Horsten and my "Earman on Underdetermination and Empirical Indistinguishability," *Erkenntnis*, IL (1998): 303-20). But that is another matter.

[2] Devitt, *Coming to Our Senses* (New York. Cambridge, 1996), p. 48.

[3] For an overview, see Ernest Sosa, "Putnam's Pragmatic Realism," this JOURNAL, XC, 12 (December 1993). 605-26.

0022-362X/99/9609/479–90

ment is the historically first—it dates back to 1976. The argument is claimed to be a reductio; it allegedly derives the negation of MF—'Some empirical theories are necessarily true'—from the conjunction of MF, CT, and SN. Although its "core" is simple and uncontested, the model-theoretic argument has continued to exercise philosophical minds ever since its first presentation, for it is not so simple to adjudicate how effective it is as an argument against realism.

Much seems to hinge here on the legitimacy of a certain crucial move in Putnam's argument. Already in his first presentation of the model-theoretic argument, he considered, and quickly dismissed, a possible objection he expected his opponents to make. Commentators have called Putnam's dismissal—the "just more theory move," as it has come to be called—"completely wrongheaded,"[4] "foul," and "illicit,"[5] and have argued that the move is "simply begging the question against the realist."[6] I aim to show here that, at least on one construal of the model-theoretic argument, Putnam's critics are wrong; on that construal, the "just more theory move" is perfectly legitimate, and the argument as a whole valid and, if we lay to rest some (what seem to be justified) doubts concerning one of its premises, even sound. There is, as far as I can see, at most circumstantial evidence for the claim that the construal of the model-theoretic argument that will be presented captures what Putnam "really had in mind." I urge, however, that the question to what extent it accords with his intentions, if at all, should not concern us too much, as I think that on the proposed reading, anyhow, his argument makes much better sense than on the more common way of interpreting it.

## I THE MODEL-THEORETIC ARGUMENT

The model-theoretic core of Putnam's argument is really quite simple, and clearly correct. Let $T_I$ be an epistemically ideal theory (that is, a theory possessing every conceivable virtue except, perhaps, truth). Being ideal, it will, among other things, be consistent. It then follows from some basic model theory that $T_I$ has models of every infinite cardinality. Choose from these some model $M$ of the same cardinality as the world, and establish a *satisfaction relation* (SAT) by mapping $M$ 1-1 onto the world. If $T_I$'s language is then interpreted according to SAT, $T_I$ comes out true of the world, that is, $T_I$ is true-on-SAT, or TRUE(SAT).

---

[4] James Van Cleve, "Semantic Supervenience and Referential Indeterminacy," this JOURNAL, LXXXIX, 7 (July 1992): 344-61, p. 349.

[5] Bob Hale and Crispin Wright, "Putnam's Model-Theoretic Argument against Metaphysical Realism," in Hale and Wright, eds., *A Companion to the Philosophy of Language* (Cambridge: Blackwell, 1997), pp. 427-57, pp. 440f.

[6] Devitt, *Realism and Truth*, p. 227. The objection that Putnam's argument ultimately begs the question has also been raised by (among many others) Anthony Brueckner, "Putnam's Model-Theoretic Argument against Metaphysical Realism," *Analysis*, XLIV (1984): 134-40, Craig Hansen, "Putnam's Indeterminacy Argument: The Skolemization of Absolutely Everything," *Philosophical Studies*, LI (1987): 77-99; and Mark Heller, "Putnam, Reference and Realism," *Midwest Studies in Philosophy*, XII (1988): 113-28.

It should be noticed that, apart from consistency, $T_I$'s idealness plays no role in this argument.[7] Hence the argument is really about consistent theories. The conclusion is restricted, however, by two assumptions that were tacitly made: first, that $T_I$ has at least one infinite model; and second, that the world is infinite. But if we agree to consider in our modal reasoning about theories only infinite worlds (so we do not examine what properties theories have in worlds that, following the assumption about the world's size, already for cardinality considerations could not be the actual one) and if, moreover, we restrict our attention to theories with at least one infinite model (note that all "real" empirical theories are of this sort), then the following can be taken to be the proper conclusion of the argument above:

(1) Any consistent empirical theory is TRUE(SAT) of any world on some satisfaction relation SAT (in fact, on indefinitely many such relations).

Prima facie, this conclusion appears to be quite harmless for the metaphysical realist. That any empirical theory is guaranteed to be TRUE(SAT) for some SAT does not seem to contradict the realist claim MF that no empirical theory is guaranteed to be *true*, where, according to the realist, a theory is true if it is true on the "intended" interpretation of the language in which it is formulated. The realist acknowledges that there is an abundance of correspondence relations between our vocabulary and reality, but he insists that only one of those is intended, only one of those is reference. So what is the point of the argument?

The question is not readily answered. I reserve my own answer for a later section, and shall here present the common interpretation of how Putnam's further moves bring the above considerations to bear against the realist. On this interpretation, his point is that, roughly put, there is no way for the metaphysical realist to discriminate between truth-on-some-interpretation and truth simpliciter. For, Putnam (on this reading) argues, from that perspective there can be but one distinguishing feature—call it *F*—between the possible interpretations of a theory's language: some make the theory come out true, others make it false. So the realist has only two options in giving substance to the notion of intendedness: he can define 'intendedness' with respect to interpretations to mean being a member of the class of *F*-interpretations (that is, of the class of interpretations that make the theory true) or alternatively as being a member of the class of non-*F*-interpretations. Since we certainly want our theories to be true, the former seems to be preferable as a definition. But then it indeed follows that any consistent empirical theory must be true, contrary to MF. (Of course, if the latter definition were chosen, it would follow that every empirical theory must be false, which the realist—as, presumably, anyone else—regards to be just as absurd as the conclusion that every such theory must be true.)

---

[7] As was already noted in David Lewis, "Putnam's Paradox," *Australasian Journal of Philosophy*, LII (1984): 221-36, which also contains a lucid discussion of the question why Putnam himself restricts the model-theoretic argument to ideal theories (pp. 230f.).

In this reasoning, two premises are crucial:

(2) If there is not some (set of) feature(s) $F^*$, other than *F*, distinguishing among all the existing correspondence relations between the words of a theory's language and (sets of) things in the world the putative unique reference-determining correspondence relation, then a theory is true exactly if it is TRUE(SAT) on some correspondence relation SAT.

(3) There is no such feature $F^*$ that complies with the metaphysical realist's standards.

Both in Putnam's presentations of the model-theoretic argument and in the responses to it of his realist opponents, (3) has received most of the attention. We shall see further on that there is an interpretation of the model-theoretic argument so that (3) is not among the argument's premises. But (2), or something close to it, must be among the premises on any interpretation. At least it is hard to see how else we are to arrive at a conclusion about the truth simpliciter of theories from the model-theoretic point concerning the truth-on-SAT (on some SAT) of theories. Putnam himself says surprisingly little about this premise.[8] Perhaps he takes it to be so obvious as to need no mention or comments. I surmise that Putnam regards (2) to be a conceptual truth, at least from the point of view of the correspondence theorist, following from the conception of truth as correspondence to reality. It indeed seems that, if there is no feature $F^*$ different from $F$, then TRUE(SAT) on some SAT is the only way to capture the idea of correspondence to reality—and under the presupposition it seems to capture that idea *fully*. One may still doubt whether the realist is committed to (2). But I shall here grant Putnam that (2) is a consequence of CT. Strictly speaking, then, my claim is a conditional one, namely, that good sense can be made of his argument provided (2) follows from CT and is thus legitimate as a premise in an argument against realism. This does not weaken my case too much, I think, given that (2), or some thesis equivalent to it, is also among the premises of the argument as commonly read, and that realist complaints about that are rare.

---

[8] I discuss this premise (or rather a principle closely related to it) in greater detail in my "A Note on Global Descriptivism and Putnam's Model-Theoretic Argument," *Australasian Journal of Philosophy* (in press). See also Mary Kate McGowan, *Realism or Non-Realism: Undecidable in Principle, Decidable in Practice* (Doctoral Dissertation, Princeton University, 1996); and van Fraassen, "Elgin on Lewis's Putnam's Paradox," this JOURNAL, XCIV, 2 (February 1997): 85-93, p. 86, and "Putnam's Paradox: Metaphysics Revamped and Evaded," *Philosophical Perspectives*, XI (1997): 17-42.

### II. JUST MORE THEORY?

(3) has stunned many of Putnam's realist opponents. Typically, naturalistically inclined realists endorse some version of the *causal theory of reference* (CTR) according to which reference is constituted causally.[9] In other words, they claim that there is an acceptable feature $F^*$—causality—sufficient to single out reference among the welter of word/world relations. The realist answer to Putnam's model-theoretic argument is then obviously this: since there is no guarantee that SAT coincides with the causal relations constituting reference, there is no guarantee that $T_I$ is true (instead of just TRUE(SAT)).

But Putnam had anticipated this response, arguing that invoking CTR cannot block the model-theoretic argument, since CTR is "just more theory." His argument for this claim could be more detailed, but it is commonly read as follows: for CTR to be a reasonable candidate theory of reference in the first place, it must be compatible with the epistemically ideal theory $T_I$ (given the constraints on $T_I$, CTR might even be assumed to be part of that theory). So CTR can be consistently added to $T_I$. But now the model-theoretic argument, being as general as it is, can simply be repeated, this time for $T_I \cup$CTR. Quite evidently—the argument on this reading continues—if we assumed that the problem the realist faces had already been solved for the part of the language in which CTR is couched, then this theory might plausibly be held to solve the problem for $T_I$ (as well as for any other theory to which the model-theoretic argument applies), but "how 'causes' can uniquely refer is as much of a puzzle as how 'cat' can, on the metaphysical realist picture."[10] In other words, CTR is part of the problem not the solution sought. To assume otherwise is to beg the issue.

---

[9] See, for example, Devitt, *Designation* (New York: Columbia, 1981) and *Coming to Our Senses*; and Clark Glymour, "Conceptual Scheming, or Confessions of a Metaphysical Realist," *Synthese*, LI (1982): 169-80.

This "just more theory move" is the most controversial part of Putnam's argument; we shall shortly discuss its legitimacy. Before doing so, however, we should briefly consider Putnam's claim that the model-theoretic argument works only against realism because SN is part of that position, that is, that there is an effective response to the argument open to an advocate of CT and MF but explicitly not SN. This claim will be seen to give independent support to my own reconstruction of the argument as opposed to the common reading of it.

The response Putnam sketches is to rely on what he calls a "magical theory of reference." Such a theory at bottom comes to a denial that reference is anything like a natural relation, that anything informative from a scientific perspective can be said about it.[11] So it could be claimed, for instance, to be a "surd, metaphysical truth"[12] that some one word/world relation is reference ("it just happens to be so"). Or some more substantive explanation of reference in terms of supernatural powers (noetic rays, intuition, direct grasping, what have you) might be tried. But how the details of such a magical theory of reference are filled in does not matter; all that matters is that SN and the standards of acceptability it imposes on theories of reference are given up.

David Lewis finds Putnam's claim that magical theories of reference can pull the trick where naturalistic ones fail puzzling. For, he thinks, if the "just more theory move" were legitimate, then it should work no matter what kind of feature *F**—whether natural or non-natural—an opponent might appeal to. As Lewis says:

> What's the good of holding up yet another sign, thus
>
> DIRECTLY GRASPS

---

[10] Putnam, *Meaning and the Moral Sciences*, p. 126. See also "Models and Reality," in his *Realism and Reason* (New York: Cambridge, 1983), pp. 1-25, pp. 17f., and "Is the Causal Structure of the Physical Itself Something Physical?" in his *Realism with a Human Face* (Cambridge: Harvard, 1990), pp. 80-95, pp. 85f.

[11] Note that, as defined here, the "account" of reference Putnam presents in his *Reason, Truth and History* (New York. Cambridge, 1981)—according to which all that can be said about reference are pragmatic tautologies, such as that 'cat' refers to cats—counts as magical, too.

[12] Putnam, *Reason, Truth and History*, p. 46.

> or perhaps
>
> INTUITS
>
> if it is still open to Putnam to challenge the determinate reference of the words written on the sign?...Putnam...has said nothing to show why the ['just more theory'] treatment applies only to natural constraints (*op. cit.*, p. 233).[13]

So if reference is claimed to be singled out by some supernatural power, why not add 'Reference is singled out by some supernatural power', and perhaps some account of that supernatural power, to $T_I$ and apply to that theory the same bit of model theory that was applied before to $T_I \cup \mathrm{CTR}$? The "unnaturalized" metaphysical realist would thereby not fare any better than the naturalized in the light of the model-theoretic argument. Although it is true that Putnam does not say much about why he thinks his argument only affects the naturalist, we shall see that, on the reconstruction of the argument to be presented, it is perfectly plausible that the distinction should be essential in Putnam's eyes.

We now turn to the evaluation of Putnam's argument, and in particular to the evaluation of the "just more theory move" therein. As intimated earlier, realists have protested vigorously against his dismissal of CTR as being just more theory and have claimed that the gambit is itself question-begging. Let us *pro temp* suppose that the common interpretation of Putnam's model-theoretic argument is correct, and that it correctly locates the place of the "just more theory move" in the argument. Then it is plain how it is that he begs the issue against the realist: to come from '$T_I \cup \mathrm{CTR}$ is TRUE(SAT)' to '$T_I \cup \mathrm{CTR}$ is true', Putnam must assume (3). But to assume (3) is to assume that CTR fails. For if CTR is true, (3) cannot be. He might answer that he does *not* assume CTR to be false. Even if CTR is true—he might continue—and states a naturalistically acceptable constraint on reference, as the realist claims, I, Putnam, have just shown the constraint to be insufficient to *determine* reference. After all, we can find any number of satisfaction relations on which $T_I \cup \mathrm{CTR}$ comes out true, can we not? But as virtually any commentator has pointed out, this would be to miss a crucial distinction, namely, that between an interpretation satisfying some *constraint* $C$ and an interpretation satisfying a *theory stating C*. Evidently, the realist never claimed that reference is fixed by the *words* of the realist theory of reference, but by the actual reference-determining mechanism the theory is a theory *of*.[14]

---

[13] Cf. also Hale and Wright, who contend that the "just more theory move" "[i]f allowable...will be available against any specific constraint the metaphysical realist may propose, regardless of its precise content ..." (p. 440).

What should make us suspicious about our understanding of Putnam's argument here is that, if the standard realist analysis of the argument is along the right lines, it would make that argument so *evidently* mistaken. It would mean that Putnam simply takes (3) for granted, whereas the dialectical situation quite obviously requires him to argue for it. Indeed, the mistake would be so obvious that it is hard to imagine a philosopher as astute and logically sophisticated as Putnam committing it.

Could it be, then, that David Anderson[15] is right that the model-theoretic argument is not the argument all or most of Putnam's critics have taken it to be, but that the full model-theoretic argument is an argument in two stages: one stage in which Putnam argues that, if (3) is the case, then metaphysical realism is false; and another stage in which he seeks to defend (3)? His argument for (3), on Anderson's reconstruction, is (in a nutshell) that, if anything could fix reference from a naturalist perspective, it is causality, but because the only notion of cause we can make sense of (physical or empirical cause) is generally overdetermined, causality will not work as a reference fixer.

It is true that at many places in his work, Putnam argues for the aforementioned claim, but, as Michael Devitt[16] rightly remarks, this is a separate argument against the realist and not in any sense part of the model-theoretic argument. Putnam's treatment of CTR, in the model-theoretic argument, has nothing to do with the specific content of CTR;[17] it has nothing to do with whatever problems might beset the notion of cause or with any other "intrinsic" shortcomings CTR might have, and this is so both on the standard interpretation of Putnam's argument and on mine below.[18] Another reason to doubt the correctness of Anderson's reading of the model-theoretic argument is that, although he is apparently aware that the argument is intended as a reductio of metaphysical realism (*op. cit.*, p. 312), on his construal it is not (or at least not obviously so): the argument for (3), as he presents it, draws on much more than is implied by the premises representing realism (whether we take these to be P1 and P2 in his paper or the premises CT, MF, and SN of our exposition).[19]

---

[14] See, for example, Glymour; Lewis; Hale and Wright; Devitt, "Realism and the Renegade Putnam," *Noûs*, XVII (1983): 291-301, *Realism and Truth*, "On Determining Reference," in Alex Burri, ed., *Language and Thought* (New York: de Gruyter, 1997), pp. 112-21; Ian Hacking, *Representing and Intervening* (New York: Cambridge, 1983); Hartry Field, "Are Our Logical and Mathematical Concepts Highly Indeterminate?" *Midwest Studies in Philosophy*, XIX (1994): 391-429; and William Alston, *A Realist Conception of Truth* (Ithaca: Cornell, 1996), p. 141. Hale and Wright suggest, and rightly reject, a further response Putnam might make at this point (p. 441).

[15] "What Is the Model-theoretic Argument," this JOURNAL, XC, 6 (June 1993): 311-22.

[16] "On Determining Reference."

[17] As we saw, this was also noted by Hale and Wright (pp. 439ff.) and Lewis.

Must we then, after all, conclude that the common reading of Putnam's model-theoretic argument is correct and thus that the argument rests on an elementary mistake? I think not. The next section offers a more charitable rendition of the argument, one on which the argument comes out perfectly valid.

## III. THE ARGUMENT RECONSTRUCTED

On my reading of Putnam's argument, realists are right that in it he assumes CTR to be false. What I would want to deny is that Putnam is not entitled to do so. Here I shall argue that there is a way of reading his argument so that the assumption that CTR is false is legitimately made. On the proposed reading, the assumption that CTR is false is made in a subargument of the full model-theoretic argument, and is discharged at the end of that subargument. Thereby we obtain a derivation of ¬MF from the uncontroversial premises MF and (1) and from the somewhat less uncontroversial (2). It is important, however, that (3) is not among the premises.

Before we go on to consider the alternative reading of Putnam's argument, a few words should be said about the argument's scope. Surely, Anderson is right in pointing out that, according to Putnam, causality is the only candidate reference fixer from a naturalistic perspective. Moreover, some realists think likewise.[20] But as I see it, Putnam's argument in no way depends on this kind of pessimism about naturalistically acceptable alternatives to causality as reference fixer.

---

[18] That Putnam at so many places in his work extensively argues against CTR itself might be interpreted as an indication that he himself is doubtful about the "just more theory move"—for if that works, it seems that any argument specifically directed against CTR is otiose. In view of Putnam's general style of philosophizing, however, it is more plausible to assume that he seeks to show that the realist runs into trouble at every turn.

[19] Moreover, if Putnam's argument should indeed be intended as Anderson claims it is, then I think Manuel García-Carpintero is right in pointing out that "the specifically model-theoretic considerations are irrelevant, and could lead to spurious debates"—"The Model-Theoretic Argument: Another Turn of the Screw," *Erkenntnis*, XLIV (1996): 305-16, p. 305

[20] See, for instance, Devitt, "Meanings Just Ain't in the Head," in George Boolos, ed., *Meaning and Method* (New York: Cambridge, 1990), pp. 79-104, p. 83, and *Coming to Our Senses*, p 160

In what follows, we read CTR as asserting that reference is fixed by causality *or by any other naturalistic constraint.*[21] From the ensuing reconstruction it will emerge that in his argument, he can even make the stronger assumption that CTR, read in this way, is false without begging any realist issues.

The basic idea underlying the reconstruction is this: Putnam seeks to establish the falsity of MF. To argue toward a reductio, he assumes the thesis to be true. Recall now that what this thesis expresses is that every scientific theory, no matter how good our evidence for it, may be false. Assuming this to be true, it follows that CTR, being an empirical theory, may be false. It is then obviously legitimate to go on: "Suppose it *is* false..." (that is, to enter a "necessitated subproof," as it is called in some proof systems for modal logic), as Putnam does, at least on the interpretation of his argument I favor. The following natural-deduction style representation of how the argument-as-I-think-of-it goes may be helpful.[22] For ease of exposition, I assume the variable $x$ to range over all and only (actual and potential) consistent empirical theories, so that we can formalize MF as $\forall x \Diamond \neg \mathrm{True}(x)$ (that is, the antecedent 'if $x$ is an empirical theory'-part of MF need not be expressed in the formalization of the thesis).

$$
{}_{\neg I}\dfrac{{}_{\neg E}\dfrac{{}_{\Box I}\dfrac{\neg\bot}{\Box\neg\bot\,(=\neg\Diamond\bot)} \qquad {}_{\Diamond E}\dfrac{{}_{\forall E}\dfrac{[\forall x\Diamond\neg\mathrm{True}(x)\,(=\mathrm{MF})]^1}{\Diamond\neg\mathrm{True}(\mathrm{CTR})} \qquad \dfrac{\begin{matrix}[\neg\mathrm{True}(\mathrm{CTR})]^2\\ D\\ \mathrm{True}(\mathrm{CTR})\end{matrix} \quad [\neg\mathrm{True}(\mathrm{CTR})]^2}{\bot}{}_{\neg E}}{\Diamond\bot}{}_{2}}{\bot}}{\neg\mathrm{MF}}{}_{1}
$$

This is only a first approximation, of course. But it does suffice to show that, by reflecting on what MF says, we might be able to get a sound derivation of the negation of MF from MF and from whatever premises may be hidden inside $D$.

---

[21] It is not clear whether CTR (however read) can be expressed in a first-order language. This might matter, since Putnam's argument, at least in its original presentation, pertains only to first-order theories. In a later presentation ("Models and Reality," p. 23), Putnam argues that the first-order/higher-order distinction is not really essential to his argument; although for higher-order languages the argument must be different, the same result can be obtained. Putnam may be too quick on this issue—cf. Hansen for some sensible comments on this point. I shall here sidestep that discussion, however, and assume that CTR is a first-order theory.

[22] The reader unfamiliar with the natural deduction system employed here may consult, for example, Neil Tennant, *Natural Logic* (Edinburgh: University Press, 1990); or Dirk van Dalen, *Logic and Structure* (New York: Springer, 1994, 3rd ed.).

In a necessitated subproof we allow, apart from the subproof's hypothesis, only hypotheses that are necessarily true in the world in which all the argument's premises are true. We so to speak enter a possible world[23] in which a certain hypothesis—in our case ¬True(CTR)—holds, in which, of course, all tautologies hold, and, most importantly, in which everything is true that according to the argument's premises is true in any possible world.[24] (1) evidently expresses a necessary truth; it can be formalized thus: $\Box\forall x\exists\mathrm{SAT}[\mathrm{TRUE(SAT)}(x)]$. (2) was granted to be a conceptual truth, at least from the perspective of the realist. If it is, then, from that perspective, it is also necessary, and can be formalized as: $\Box((3) \rightarrow \forall x(\exists\mathrm{SAT}[\mathrm{TRUE(SAT)}(x)] \leftrightarrow \mathrm{True}(x)))$. Finally, it must be noted that (3) is equivalent to the assumption of CTR's falsity (CTR understood in the "broad" sense defined earlier).[25] Thus, $\neg\mathrm{True(CTR)}\leftrightarrow(3)$ is a tautology.

We are now ready to state the subproof:

$$
{}_{\Diamond E}\dfrac{\Diamond\neg\mathrm{True(CTR)} \qquad {}_{\to E}\dfrac{{}^{1}[\neg\mathrm{True(CTR)}] \qquad {}_{\to E}\dfrac{{}_{\forall E}\dfrac{{}_{\wedge E}\dfrac{{}_{\to E}\dfrac{{}_{\to E}\dfrac{{}^{1}[\neg\mathrm{True(CTR)}] \quad \dfrac{\neg\mathrm{True(CTR)} \leftrightarrow (3)}{\neg\mathrm{True(CTR)} \rightarrow (3)}{}_{\wedge E}}{(3)} \quad (3)\rightarrow\forall x(\exists\mathrm{SAT}[\mathrm{TRUE(SAT)}(x)] \leftrightarrow \mathrm{True}(x))}{\forall x(\exists\mathrm{SAT}[\mathrm{TRUE(SAT)}(x)] \leftrightarrow \mathrm{True}(x))}}{\forall x(\exists\mathrm{SAT}[\mathrm{TRUE(SAT)}(x)] \rightarrow \mathrm{True}(x))}}{\exists\mathrm{SAT}[\mathrm{TRUE(SAT)}(\mathrm{CTR})] \rightarrow \mathrm{True(CTR)}} \quad \dfrac{\forall x\exists\mathrm{SAT}[\mathrm{TRUE(SAT)}(x)]}{\exists\mathrm{SAT}[\mathrm{TRUE(SAT)}(\mathrm{CTR})]}{}_{\forall E}}{\mathrm{True(CTR)}}}{\bot_1}}{\Diamond\bot}
$$

What we get when we fit this into the earlier schema is a refutation of metaphysical realism which only depends on premises acceptable to

---

[23] This should not be taken too literally, if only because the assumption that there is a possible world of a certain kind may lead to a contradiction so that there can be nothing we might have entered in our subproof.

[24] So the ◇E-rule can alternatively be stated as follows:

$$
\dfrac{\Box\Gamma,\ \Diamond\varphi \qquad \begin{matrix}[\varphi]^1\ [\gamma_1]^2 \dots [\gamma_n]^{n+1}\\ \vdots \\ \psi\end{matrix}}{\Diamond\psi}{}_{1,\ ,n+1}
$$

where $\Box\Gamma$ abbreviates $\Box\gamma_1 \dots \Box\gamma_n$. It is easily seen that for ◇E to be sound no requirements regarding the accessibility relation have to be made, which means that it can serve as a rule in a natural deduction presentation of any modal logic. This also holds true for the other modal rule involved in the reconstruction.

[25] The same holds, from Devitt's viewpoint, for CTR understood narrowly. According to him, causality is the only naturalistically acceptable candidate reference fixer, so if CTR (narrowly understood) is false, (3) must be true, and vice versa.

the realist. In particular, (3) is not among the argument's premises, as promised. Furthermore, the reconstruction immediately suggests an answer to Lewis's puzzle why Putnam thinks supernatural constraints are exempt from the "just more theory treatment." Consider the non-naturalist realist who endorses some magical, nonscientific, theory of reference MTR. From the fact that he avows MF, it does not follow that he is committed to 'Possibly, MTR is false', for MF is a thesis about *scientific* theories; it says nothing about magical theories. Hence, an account of some supernatural power, or even just the claim that reference is fixed by supernatural means, though a theory in the strict sense of that word (it *is* a set of sentences, of course), is not just more theory in the sense which makes it susceptible to Putnam's model-theoretic considerations—it is not just more *empirical* theory. It is indeed the naturalist's insistence that semantics, properly pursued, is a branch of natural science that plays havoc on him; if a theory of reference is to be an empirical theory, then it is automatically within the scope of MF.

Thus, the reading of Putnam's argument offered here shows the argument to be valid, contrary to what most if not all realists believe, and would in fact rightly believe if the common reading were the only one possible. As a further advantage, the alternative reading solves Lewis's puzzle why Putnam thinks his argument only holds against the semantic naturalist. If my reading should not be the 'intended one', then, I think, the former constitute good reasons why Putnam *should* have intended it.

IGOR DOUVEN

Utrecht University

THE JOURNAL OF PHILOSOPHY
VOLUME XCVI, NO. 10, OCTOBER 1999

UNIVER JAIPUR CAMPUS

# IS THERE A PLACE FOR PHILOSOPHY IN QUINE'S THEORY?*

In the early part of the twentieth century, the logical positivists launched a powerful attack on traditional philosophy, rejecting the very idea of philosophy as a substantive discipline and replacing it with a practical, conventionalist, metatheoretical view of philosophy. The positivist critique was based on a series of dichotomies: the analytic versus the synthetic, the external versus the internal, the apriori versus the empirical, the metatheoretical versus the object-theoretical, the conventional versus the factual. W. V. Quine's attack on the positivists' dichotomies was, by extension (if not by intention), also an attack on their critique of philosophy. In time, however, his own theory took an extreme naturalistic turn which, if anything, deepened the schism between philosophy and knowledge.

I would like to show that many of Quine's early philosophical ideas—the denial of the analytic-synthetic distinction, the theses of the interconnectedness of knowledge, universal revisability, and inseparability of language from theory, his methodological pragmatism and realism—are in fact compatible with a substantive philosophy. Moreover, certain inner tensions in Quine's theory naturally lead to a new model of knowledge in which philosophy plays a substantive role: not as a "first philosophy," or as a "metascience," or as a "chapter in psychology," but as an independent discipline in its

---

* An earlier, abridged version of this article, entitled "On the Place of Philosophy in Quine's Early Theory," appears in P. Weingartner, G. Schurz, and G. Dorn, eds., *The Role of Pragmatics in Contemporary Philosophy: Papers of the 20th International Wittgenstein Symposium*, Volume II (Kirchberg am Wechsel: Austrian Wittgenstein Society, 1997), pp. 889-95.

0022-362X/99/9610/491–524

own right, alongside, continuous with, and complementary to, science.[1]

## I. AN INNER CONFLICT

In his 1973 paper, "The Significance of Quine's Indeterminacy Thesis," Michael Dummett[2] makes the following provocative claim: Quine rejected the analytic-synthetic distinction in "Two Dogmas of Empiricism" (1951) on the ground that no coherent and philosophically significant content has ever been assigned to it. Yet, ironically, the distinction has come to life for the first time in the very account that was meant to replace it: Quine's holistic model of knowledge with its center-periphery structure. Dummett's claim is based on two observations: (1) Quine's center-periphery model provides us with notions in terms of which the analytic-synthetic distinction is well defined. Roughly, analytic statements are immune to experience, synthetic statements are vulnerable to experience (that is, their revision would be mandated by suitable experiences).[3] (2) This version of the distinction cannot be rejected as empty, since the claim that there are no synthetic statements and no analytic statements (in the above sense) conflicts with the center-periphery model. Conclusion: Quine's claim that the analytic-synthetic distinction is unfounded (henceforth, the *negative analytic-synthetic thesis* or NAS) conflicts with his claim that our body of knowledge has a center-periphery structure (henceforth, the *center-periphery thesis* or CP[4]). I shall not discuss Dummett's account of the conflict here: for one thing, my point of view is somewhat different from his; for another, I have some difficulty with certain aspects of his argument. Instead, I shall present the conflict in my own way and from my own perspective, namely, as a conflict between two epistemic methodologies: the one associated with NAS, the other with CP.[5]

---

[1] For the positivists' view of philosophy, see Rudolf Carnap, *The Logical Structure of the World and Pseudoproblems in Philosophy*, R.A. George, trans. (Berkeley. California UP, 1967), *The Logical Syntax of Language*, A. Smeaton, trans. (London: Kegan Paul, 1937); *Foundations of Logic and Mathematics, International Encyclopedia of Unified Science*, Volume I, Number 3 (Chicago: University Press, 1939), "Empiricism, Semantics, and Ontology," in *Meaning and Necessity: A Study in Semantics and Modal Logic* (Chicago: University Press, 1956, second edition), pp. 205-21; A.J. Ayer, *Language, Truth and Logic* (London: Gollancz, 1946, second edition). By 'Quine's early ideas' I mean Quine's views in "Two Dogmas of Empiricism," in *From a Logical Point of View* (Cambridge: Harvard, 1980, second edition), pp. 20-46; "Truth by Convention," in *The Ways of Paradox and Other Essays* (Cambridge: Harvard, 1976, rev ed.), pp. 77-106; "Carnap and Logical Truth," in *The Philosophy of Rudolf Carnap*, P.A Schilpp, ed (La Salle: Open Court, 1963), pp. 385-406; and other articles from the same period. I am not the first to suggest that some of Quine's ideas point in the direction of a substantive philosophy. See, for example, Paul Gochet, *Ascent to Truth. A Critical Examination of Quine's Philosophy* (Munich: Philosophia, 1986), p. 16. See also George D. Romanos's analysis of Quine's conception of philosophy in *Quine and Analytic Philosophy: The Language of Language* (Cambridge: MIT, 1983), pp. 187-95. Romanos's analysis, however, is quite different from mine.

[2] In *Truth and Other Enigmas* (Cambridge: Harvard, 1978), pp. 375-416.

[3] In Dummett's words: "an analytic sentence is one such that no recalcitrant experience would lead us to withdraw our assignment to it of the value true, while a synthetic one is one such that any adequate revision prompted by certain recalcitrant experiences would involve our withdrawing an assignment to it of the value true" (*ibid.*, p 375).

A methodological conflict. The most basic dichotomy underlying the analytic-synthetic distinction is that of statements grounded in *matters of fact* and statements grounded in something *other than fact*. What this other thing is varies from one philosophical theory to another, but Quine's attack applies equally to older as well as to newer renderings of the distinction. Methodologically, the most important feature of the analytic-synthetic thesis is the *bifurcation of our standards of knowledge*: the standards for the acceptance, justification, revision of synthetic items of knowledge are different from the standards for the acceptance, justification, revision of analytic items. Quine's revolution consists in rejecting this bifurcation: every item of knowledge is subject to both standards—logic may be revised in response to experience, observation statements may be preserved based on pragmatic considerations.

Now, CP breaks this uniform picture of knowledge by positing a structure with two essentially different, if interconnected, zones: an inner zone—*center*—based on one set of standards, and an outer zone—*periphery*—based on another. Elements in the center are manipulated using pragmatic standards, elements in the periphery using evidential (experiential) standards. Elements located in the periphery stand in a privileged relation to reality from which elements located in the center are excluded. Thus, granted that a conflict with experience can be resolved by changes anywhere in the system, there is still an essential difference between changes occurring in the center and changes occurring in the periphery. The ground for changing statements in the periphery is the need to square these statements themselves with experience; the ground for changing statements in the center is the need to square other statements, namely, statements in the periphery, with experience. Truths located in the center are held solely in virtue of their contribution to the overall working of the system, hence due to pragmatic considerations. Truths located in the periphery are held partly because of their contribution to the system as a whole, but largely, and most importantly, due to their own direct link to experience, that is, based on factual, objective considerations. Take logic, for example. There is no room in Quine's model for a conflict between a logical statement, *l*, and an external event, *e*, analogous to, say, the conflict between the peripheral statement, 'It rained in San Diego on Monday, March 1st, 1999', and the event of sunny skies in San Diego on that day. Reality never impinges as directly on logic as it does on experimental science, and logic, therefore, never occupies the periphery. We may change our logic in response to conflicts with experience, but only pragmatically and via a chain of connections starting and ending in the periphery. The bone of contention between NAS and CP is the role of fact and convention in knowledge: while a principled division of knowledge into conventional and factual constituents is denied by NAS, such a division is affirmed by CP. By superimposing a center-periphery structure on the unified domain of knowledge postulated by NAS, Quine has de facto incorporated certain significant elements of the analytic-synthetic dichotomy into his epistemology.

---

[4] I shall use 'CP' both for the "center-periphery thesis" and for the "center-periphery model."

[5] In addition to "The Significance of Quine's Indeterminacy Thesis," Dummett also discusses the conflict in *Frege. Philosophy of Language* (New York: Harper, 1973), chapter 17. By offering my own version of the conflict, I do not wish to diminish my debt to Dummett; needless to say, he is not responsible for my version.

## II. OBJECTIONS

Before turning to possible solutions, I would like to consider three likely objections to the reality of the conflict.[6]

*A. Objection one.* CP cannot be said to reproduce the analytic-synthetic distinction, since this distinction requires an absolute division of statements into two mutually disjoint groups, but the center-periphery model represents our body of knowledge as a continuum. The center-periphery duality is gradual, while the analytic-synthetic dichotomy requires sharp boundaries.

*Response.* First, I would like to emphasize that on my construal, the center-periphery duality is not said to reproduce the analytic-synthetic distinction in its entirety. At issue is a significant overlap of the two distinctions, not a complete identity (or inclusion).

Second, the claim that the analytic-synthetic distinction does not allow a graduated progression may be true to its letter but is not true of its use, especially at the hands of the logical positivists. Explanation: traditionally, the analytic-synthetic distinction is a distinction between *whole* statements; but for the logical positivists, a more basic distinction exists between *constituents* of statements. This secondary distinction plays an especially important role in Rudolf Carnap's reductionist project in *The Logical Structure of the World* (*op. cit.*). Science, in Carnap's construction, is built out of basic elements of two types: purely experiential elements (in Carnap's construction, "elementary experiences" and the primitive relation of "recollection of similarity"), and purely conventional elements (logico-mathematical operations and the linguistic operation of definition). The construction proceeds in a hierarchy of stages: in stage one, purely conventional operators are applied to purely experiential elements, generating mixed elements of level one; in stage two, conventional operators (and mixed operators, if any, obtained in stage one) are applied to elements of level one, generating mixed elements of level two, and so on. We progress in this way from elementary experiences to pairs of elementary experiences, equivalence classes of elementary experiences (under the relation of recollection of similarity relative to a given agent), and so forth. Somewhere along the way we obtain qualities (quality structures), observable physical objects, theoretical objects, and so on. The result is a continuum of elements and, corresponding to it, a continuum of statements. If we order these statements according to their balance of experiential to conventional constituents (based on some criterion or another), we arrive at a progression of statements stretching from the purely synthetic to the purely analytic. We could, of course, draw a line (mark a region) somewhere along the continuum: statements on one side of the line (inside the designated region) would be "absolutely" analytic, all the rest "absolutely" synthetic; but such a division would still, at bottom, be graduated.

---

[6] Of course, there is no limit to the number of objections that can be raised. I have selected three objections which, in addition to their intrinsic merits, provide me with the opportunity to elucidate relevant issues.

Consider Carnap's analysis of a simple synthetic statement, say, 'All dogs are carnivores'. This statement is doubly mixed: on the one hand, it contains both analytic and synthetic constituents (the analytic operators 'all' and 'if...then', and the synthetic predicates 'dog' and 'carnivore'); on the other hand, its synthetic constituents themselves are decomposed into analytic and synthetic subconstituents. The concept "dog," for example, is constructed by successive applications of analytic operations, first, to purely synthetic elements and, then, to mixed elements, generating a gradual progression of concepts (and, correspondingly, objects) from the purely synthetic to the increasingly analytic. Speaking objectually, this progression includes:

(1) Elementary experiences and the relation of recollection of similarity between them (primitive).
(2) "Dog" sense qualities: for example, the color brown (equivalence classes of elementary experiences under the relation of recollection of similarity).
(3) "Dog" sense-qualities-at-space-time-locations· for example, brown in the kennel at $t$ ($n$-tuples whose first four elements are real numbers—representing space-time locations—followed by one or more "dog" sense qualities).
(4) "Dog" points of the perceptual world (classes of "dog" sense-qualities-at-a-location).
(5) "Dog" States (classes of dog perceptual points).
(6) Individual dogs (classes of "dog" states).
(7) The general object or species *Dog* (the class of all individual dogs).[7]

What, then, is the conflict between Carnap's *graduated progression* and Quine's NAS? The conflict consists of at least three disagreements. (a) According to Carnap, our body of knowledge is ultimately built out of purely experiential and purely conventional elements; according to Quine, no constituent of knowledge is either purely experiential or purely conventional. (b) According to Carnap, each element is divided into constituents whose experiential/conventional status is uniquely determined; according to Quine, no element is uniquely divisible in this way. (c) According to Carnap, the application and satisfaction conditions of some elements are purely conventional while those of others purely factual (experiential); according to Quine, the application and satisfaction conditions of all elements are both factual and conventional.

*B. Objection two.* The center-periphery methodology differs from the analytic-synthetic methodology with respect to what is arguably the most important principle of Quine's theory: the principle of universal revisability. The analytic-synthetic methodology views analytic statements as immune to revision, synthetic (empirical) statements as linked to a range of experiences whose occurrence would (under appropriate conditions) compel their revision. In contrast, the center-periphery methodology (like the NAS methodology) views no statement as immune to revision and the revision of no statement as compelled by any (specific) experiences. Given this major gap between the two methodologies, any similarity between them is of minor significance.

---

[7] See *The Logical Structure of the World*, pp. 247-48 A more detailed account of the constructional stages of physical objects (and their concepts) appears on pp. 175-211

*Response.* The claim that the analytic-synthetic distinction is tied to the traditional revision principle is simply incorrect.[8] This claim is refuted if not by the classical, Kantian model, then by the positivist model. In this model, analytic statements are based on linguistic conventions, and these are all too easy to replace. Given an infirming experience (or a series of infirming experiences), the positivist model allows a wide array of reactions: (a) no revision; (b) revision of synthetic statements: (i) directly implicated synthetic statements (observational or "protocol" statements), (ii) indirectly implicated synthetic statements (theoretical statements); and (c) revision of analytic statements (logico-linguistic conventions and/or framework principles).

Thus, discussing a case in which a protocol sentence deduced from a physical theory is disconfirmed by experiment, Carnap says:

> There are no established rules for the kind of change which must be made. For instance, the P[hysical]-rules can be altered in such a way that those particular primitive sentences are no longer valid; or the protocol-sentence can be taken as being non-valid; or again the L[ogical-linguistic]-rules which have been used in the deduction can also be changed.[9]

Elsewhere, Carnap describes a hypothetical situation in which an empirical discovery would naturally lead to a conventional revision:

> If certain events allegedly observed in spiritualistic séances, e.g., a ball moving out of a sealed box, were confirmed beyond any reasonable doubt, it might seem advisable to use [the conventional device of] four spatial coordinates [instead of the customary three].[10]

But a conflict with experience, Carnap says, is also compatible with no revision:

> [W]hen [a scientific hypothesis] proves to be L-incompatible with certain protocol-sentences, there always exists the possibility of maintaining the hypothesis and renouncing acknowledgement of the

---

[8] Quine himself appears to endorse this claim (see "Two Dogmas," p. 43). He is wrong, in my view, unless he has in mind a restricted version of the claim (restricted, for example, to Immanuel Kant's version of the distinction or to the positivist version from a purely internal perspective)

[9] *The Logical Syntax of Language*, p. 317. (The two sentences are transposed.)

[10] "Empiricism, Semantics, and Ontology," pp 212-13

protocol-sentences.[11]

Finally, Quine's principle of universal revisability is explicitly endorsed by Carnap:

> [In "Two Dogmas"] Quine shows...that a scientist, who discovers a conflict between his observations and his theory and who is therefore compelled to make a readjustment somewhere in the total system of science, has much latitude with respect to the place where a change is to be made. In this procedure, no statement is immune to revision, not even the statements of logic and of mathematics.... With...this I am entirely in agreement.[12]

Where Carnap's liberal policy clashes with Quine's is in its strict bifurcation of the standards for revision: purely pragmatic (practical) standards for the revision of logico-mathematical statements; largely factual (theoretical, evidential) standards for the revision of physical statements.[13]

C. *Objection three.* There is no real conflict between CP and NAS, since the two theses operate on different levels. While CP is a descriptive, *behaviorist* thesis, NAS is a *prescriptive, normative* thesis. CP describes scientists' habitual practices, including their tendency to regard certain statements as conventional, others as factual and still others as occupying intermediate positions between the two extremes; NAS says that scientists' habitual practices are not inviolable principles of rational inquiry. Clearly, there is no conflict between the claim that scientists *normally act* in accordance with the analytic-synthetic distinction and the claim that scientists are not *normatively bound* by this behavior.

*Response.* This is an important objection whose validity depends on the role assigned to CP in Quine's argument. To identify this role, I propose the following analysis of the general structure of "Two Dogmas of Empiricism." The paper is divided into two parts: sections one to five, and section six, respectively. The first part consists of a series of arguments to the effect that (a) the analytic-synthetic distinction is not well defined, and (b) the analytic-synthetic distinction is connected with a failed epistemology (namely, reductionism).[14] None of these arguments conclusively demonstrates the unacceptability of the distinction. The first argument (cluster of arguments) merely shows that past attempts to define the distinction have failed, not that any attempt to define it is bound to fail; the second argument shows that a popular epistemology incorporating the distinction suffers serious difficulties, not that any epistemology incorporating it is bound to suffer these difficulties, or that these difficulties are in principle insurmountable. In spite of the inconclusive nature of these arguments, Quine arrives at a radical and uncompromising conclusion: the analytic-synthetic distinction should be altogether and in principle banished from philosophy.

---

[11] *The Logical Syntax of Language*, p. 318.

[12] "Replies and Systematic Expositions," in *The Philosophy of Rudolf Carnap*, pp. 859-1013, here p. 921.

[13] For an especially sharp version of this bifurcation, see Carnap's "Empiricism, Semantics, and Ontology."

Quine's "jump" from the claim that past attempts to establish the distinction face serious difficulties to the claim that the distinction is (in principle) an untenable dogma is obviously not a step in deductive reasoning. Rather, it is a methodological change of direction, similar to other methodological shifts in the history of philosophy: for example, Immanuel Kant's so-called "Copernican" revolution. We could, indeed, use roughly the same reasoning as Kant's (or the same skeleton of reasoning) to close the gap in Quine's argument:

> Hitherto it has been assumed that all our knowledge must be reducible to ultimate elements of two kinds—analytic and synthetic. But all our attempts to provide a solid foundation for knowledge based on this division have ended in failure. We must therefore make trial whether we may not have more success in the task of epistemology, if we suppose that knowledge is not divisible in this way.[15]

Quine's rejection of the analytic-synthetic dichotomy, like Kant's rejection of the traditional world-observer dichotomy, is a promissory note, to be made good by a new model of knowledge which demonstrates the virtues of the new methodology. The construction of such a model is undertaken in the second part (section six) of "Two Dogmas," where Quine draws an outline of a new epistemic model, CP, whose task is to vindicate NAS. Quine is quite clear about the intended role of CP. Immediately following the introduction of the model, he announces its direct implicational relation to NAS:

---

[14] For exposition and critical discussion of these arguments, see Gochet; H.P. Grice and P.F. Strawson, "In Defence of a Dogma," *The Philosophical Review*, LXV (1956): 141-58; Roger F. Gibson, *The Philosophy of W.V. Quine: An Expository Essay* (Tampa: South Florida UP, 1982); Paul A. Boghossian, "Analyticity Reconsidered," *Noûs*, XXX (1996): 360-91.

[15] An adaptation of a passage in *Critique of Pure Reason* (1787, second edition): "Hitherto it has been assumed that all our knowledge must conform to objects. But all attempts to extend our knowledge of objects by establishing something in regard to them *a priori*, by means of concepts, have, on this assumption, ended in failure. We must therefore make trial whether we may not have more success in the tasks of metaphysics, if we suppose that objects must conform to our knowledge" (Bxvi)—Norman Kemp Smith, trans. (London: Macmillan, 1929).

> If this view [the view of knowledge represented by CP] is right,...it becomes folly to seek a boundary between synthetic statements...and analytic statements....[16]

Furthermore, tenets central to NAS follow from the model:

> ...it is misleading to speak of the empirical content of an individual statement.... Any statement can be held true come what may, if we make drastic enough adjustments elsewhere in the system.... Conversely, by the same token, no statement is immune to revision... (*ibid.*).[17]

It is evident that Quine regards CP as a *realization* of NAS. The model is designed to justify NAS by example, not by counterexample. CP and NAS are not two independent theses, the one descriptive, the other normative, the one capturing one aspect of our system of knowledge, the other, another. Rather, the two theses are designed to complement one another: CP by providing an attractive picture of knowledge in accordance with NAS, NAS by motivating and articulating the normative principles underlying CP. Since NAS is a normative thesis, CP is assigned the role of a normative model exemplifying its principles.[18]

If this analysis is correct, the third objection is dissolved: CP is a normative thesis and its conflict with NAS is a genuine conflict, threatening the integrity of Quine's project. In the next section, I shall consider, and reject, two eliminative solutions to this conflict. In the last two sections, I shall present a third, constructive solution, with substantial ramifications for philosophy.

## III. TWO SOLUTIONS

I shall begin with two eliminative solutions: (*A*) *withdraw NAS,* (*B*) *withdraw CP.*

---

[16] "Two Dogmas of Empiricism," p. 43.

[17] The revisability principle, as we have seen above, is not unique to NAS, but it still constitutes an important component of NAS.

[18] One may try to argue for a descriptive reading of CP based on the third paragraph of section 6: "For vividness I have been speaking..." (*op. cit.*, pp. 43-44). But even if it is natural to read a few sentences in this paragraph as suggesting a descriptive model, my argument is not affected. What I have attempted to do is *not* to show that no sentence in "Two Dogmas" is compatible with a descriptive model. Rather, I have attempted to show that (i) a descriptive model of knowledge would be extraneous to Quine's main argument in "Two Dogmas," (ii) the main argument of "Two Dogmas" calls for a normative model, and (iii) there is strong textual evidence to support a normative construal of the model. In contrast, interpreting CP as a descriptive model would raise difficult questions with respect to the interpretation of Quine's overall argument, the contribution of CP to this argument, and the role of the indicated passage in an otherwise normative section, sandwiched between two clearly normative passages. My tendency is to view the third paragraph as a digression, possibly foreshadowing Quine's later epistemology.

*A. First solution: withdraw NAS.* This solution is recommended by Dummett. Quine's main achievement in "Two Dogmas of Empiricism," according to Dummett, is a new verificationist model of language (knowledge) which, by being moderately holistic ('organic', in Dummett's terminology), avoids the pitfalls of earlier models. Since NAS threatens to undermine this model, Dummett recommends the elimination of NAS. He finds support for his proposal in Quine's own "change of heart" in *Word and Object*[19] and later writings, where, according to Dummett, Quine has "quietly dropped" NAS.[20]

*Disputation.* I shall offer three reasons—exegetical, historical, and substantive—for rejecting this solution. The second reason I shall merely allude to, the other two I shall discuss in more detail.

*(1) Exegesis. (a) Early writings.* It is clear that prior to *Word and Object*, Quine's espousal of NAS was full hearted and uncompromising. Note the sharp tone and disparaging expressions in the following citations from "Two Dogmas": "for all its a priori reasonableness, a boundary between analytic and synthetic statements simply has not been drawn. That there is such a distinction to be drawn at all is an unempirical dogma of empiricists, a metaphysical article of faith"; "it is nonsense, and the root of much nonsense, to speak of a linguistic component and a factual component in the truth of any individual statement"; "If this view is right,...it becomes folly to seek a boundary between synthetic statements...and analytic statements" (*op. cit.*, pp. 37, 42, 43). In other articles from the same period, Quine's message is just as strong: "My misgivings over the notion of analyticity are... misgivings in principle"; "analyticity...is a pseudo-concept which philosophy would be better off without."[21] In contrast, Quine's advocacy of the center-periphery duality is somewhat qualified: the idea of a distance from a center and a periphery is merely a "metaphor," brought in for the sake of "vividness" and not to be taken quite literally.[22] There is no doubt that faced with a decision between the two theses, Quine of "Two Dogmas" would have chosen NAS.

*(b) Later writings.* In spite of endorsing a few watered-down versions of the analytic-synthetic distinction in some of his later writings,

---

[19] Cambridge: MIT, 1960.

[20] See Dummett, "The Significance of Quine's Indeterminacy Thesis," pp. 377-78.

[21] "Mr. Strawson on Logical Theory" and "Three Grades of Modal Involvement" in *The Ways of Paradox and Other Essays*, pp. 137-57 and pp. 158-76, here p. 139 and p. 171, respectively.

[22] See "Two Dogmas," p. 43.

Quine has never wavered in his rejection of the full-blooded distinction. Characterizing his own version in *Word and Object* (stimulus analytic versus stimulus synthetic) as a "behavioristic ersatz" and a "vegetarian imitation" (*op. cit.*, pp. 66, 67), Quine declares that the full-scale distinction "only encourages confused impressions of how language relates to the world" (*op. cit.*, p. 67). Then, commenting on the customary grouping of "synthetic truth," "fact," "objectivity," and "accessibility to observation," he says: "So here that same analytic-synthetic dichotomy intrudes which we have found so dubious" (*op. cit.*, p. 247). And as usual, Quine is highly critical of Carnap's uses of the distinction: "Carnap has long held that the questions of philosophy, when real at all, are questions of language.... But why should this be true of the philosophical questions and not of theoretical questions generally? Such a distinction of status is of a piece with the notion of analyticity..., and as little to be trusted" (*op. cit.*, p. 271). Quine's impatience with "analyticity" is once again revealed in *Philosophy of Logic*[23]: "it is time to rein in our verbalism. What are we trying to get at when we call a sentence analytic, or true purely by virtue of the language" (*ibid.*, p. 96)? In *The Roots of Reference*,[24] Quine continues to emphasize the difference between his own version of the distinction and Carnap's: "we have here no such radical cleavage between analytic and synthetic sentences as was called for by Carnap and other epistemologists" (*ibid.*, p. 80). And as late as *Pursuit of Truth*,[25] Quine says: "The importance of analyticity for epistemology lapses, be it noted, in the light of holism" (*ibid.*, pp. 55-56). There is no question that Quine's commitment to NAS has persisted through the years.

(2) *Historical perspective.* Historically, the analytic-synthetic distinction is tied to a long chain of philosophical dichotomies: the apriori versus the aposteriori, the necessary versus the contingent, reason versus experience, reality versus appearance, objectivity versus subjectivity, knowledge versus opinion, and so on. It is hard to exaggerate the role of these dichotomies in initiating philosophical movements: rationalism, skepticism, idealism, empiricism, pragmatism, and positivism have all been driven by some or all of these dichotomies. In repudiating the analytic-synthetic distinction, NAS calls into question this entire line of philosophical distinctions.

---

[23] Englewood Cliffs: Prentice Hall, 1970.

[24] La Salle: Open Court, 1973.

[25] Cambridge: Harvard, 1990.

From a historical perspective, therefore, NAS is a revolutionary thesis, challenging some of the most fundamental assumptions of philosophy and offering an altogether new approach to basic philosophical problems. In contrast, CP treads on well-known grounds, being a variant of the logical positivist account of knowledge and an offspring of classical empiricism. From the point of view of their critical as well as their innovative potential, NAS is unquestionably of far greater philosophical interest than CP.

(3) *Strength of thesis.* I shall present an outline of three epistemic arguments in support of NAS.

(a) *A sound methodological principle.* One of the consequences of the analytic-synthetic dichotomy is a rigid methodology of confirmation and infirmation: synthetic truth is a matter of fact and is therefore confirmed or infirmed by whatever counts as factual evidence with respect to it; analytic truth is a matter of something other than fact, hence there is no such thing as factual evidence for (or against) it. Synthetic statements are justified by appeal to the world; analytic statements by appeal to *X*, where *X* is language, method, convention, thought, mind, and so on. *Synthetic truths are vulnerable to experience, analytic truths are immune to experience.* This methodological principle has never been adequately established. Indeed, one of the lessons that history has taught us is that no statement is immune to experience, and many revisions blur the line between the factual and the pragmatic: "Revision even of the logical law of excluded middle has been proposed as a means of simplifying quantum mechanics; and what difference is there in principle between such a shift and the shift whereby Kepler superseded Ptolemy, or Einstein Newton, or Darwin Aristotle?"[26]

Hilary Putnam's[27] analyses of three episodes in the history of science—the replacement of Newton's definition of kinetic energy by Einstein's, the passage from Euclidean to non-Euclidean physical geometry, and the development of a nonstandard logic for quantum mechanics—further support Quine's claim. Putnam concentrates on the factual element in the revision of "analytic" items. The revision of definitions, logical laws, and mathematical theories, according to Putnam, does not have to be conducted on a separate level of inquiry (presumably, the level of metatheory) or guided by a special set of norms (presumably, purely pragmatic norms). Rather, in the course of scientific change, a definition may come to be treated as an identity statement (a law of nature),[28] an abstract geometry as a physical theory (a theory of the geometrical features of physical space), and a logic as a method of reasoning for a particular formal structure (a structure of objects possessing certain formal characteristics). In this way, the laws of logic are, in principle, vulnerable to the world: if the formal behavior of objects conflicts (in some "deep" way) with a particular logic, then this logic is challenged by the behavior of those objects.[29]

---

[26] "Two Dogmas of Empiricism," p. 43.

[27] "The Analytic and the Synthetic," in *Mind, Language and Reality. Philosophical Papers*, Volume 2 (New York: Cambridge, 1975), pp. 33-69, and "The Logic of Quantum Mechanics" in *Mathematics, Matter and Method: Philosophical Papers*, Volume 1 (New York. Cambridge, 1979, second edition), pp. 174-97.

Using the metaphor of a battle between *nature* and *humanity* to represent the pursuit of knowledge,[30] we can describe the methodological rationale for NAS as follows: the analytic-synthetic dichotomy creates a false line of defense against nature. Nature, the analytic-synthetic dichotomy induces us to believe, is in principle incapable of threatening the analytic zone of our knowledge. But nature might (and some say, does) find ways to encroach upon this zone. The analytic-synthetic policy of complacency in the analytic zone, careful measures for establishing the correctness of our theories in the synthetic zone, is therefore unwarranted. We do not know in advance where nature will choose to strike next, and by restricting our defenses to the synthetic front of knowledge, are we not creating an epistemic Maginot line? As a matter of prudent strategy, it is incumbent upon us to maximize the maneuverability of our cognitive resources, and just this is accomplished by NAS.

---

[28] Putnam uses Albert Einstein's definition of kinetic energy to exemplify the similarity between definitions and laws. The definition of kinetic energy, according to Putnam, is motivated by the same kind of considerations and has the same kind of role in Einstein's theory as such bona fide laws as 'Moving clocks slow down' and 'Nothing exceeds the speed of light'—"The Analytic and the Synthetic," pp. 42-46. The treatment of definitions as identity statements is also familiar from logical semantics. If $T$ and $T'$ are theories formulated within the framework of standard logic, and $T$ differs from $T'$ only in having '$a =_{Df} f(b)$' where $T'$ has '$a = f(b)$' (or in having '$Ax =_{Df} Bx$' where $T'$ has '$(\forall x)(Ax \equiv Bx)$'), then $T$ and $T'$ have exactly the same models, exactly the same logical consequences, and so on, that is, $T$ and $T'$ are model-theoretically indistinguishable.

[29] This way of looking at logic is common among quantum logicians. The rules of classical logic, these logicians argue, are "obeyed" by systems of objects embeddable in a Boolean structure, but systems exhibiting other formal structures, for example, the structure of a nondistributive complemented lattice, require a different logic. See Putnam, "The Logic of Quantum Mechanics"; David Finkelstein, "Matter, Space and Logic," in Robert S Cohen and Marx W Wartofsky, eds., *Boston Studies in the Philosophy of Science*, Volume 5 (Boston. Reidel, 1969), pp. 199-215; and R.I.G. Hughes, *The Structure and Interpretation of Quantum Mechanics* (Cambridge. Harvard, 1989), chapter 7. This way of looking at logic may also underlie Alfred Tarski's statement that "certain ..experiences of a very fundamental nature may make us inclined to change just some axioms of logic"—Morton White, "A Philosophical Letter of Alfred Tarski," this JOURNAL, LXXXIV, 1 (January 1987): 28-32, here pp. 31-32.

[30] The metaphor is taken from Jaakko Hintikka; see, for example, his "Language-Games for Quantifiers," in *Logic, Language-Games and Information: Kantian Themes in the Philosophy of Logic* (New York: Oxford, 1973), pp 53-82.

One possible objection to this line of reasoning can be formulated as follows: granting the unpredictability of nature, it is still not clear how nature can surprise us when it comes to our own concepts. If the concept of *A* includes the concept of *B*, what can nature do about it? How can nature falsify our judgment that every *A* is a *B*? This challenge can be expressed both in Kantian and in positivist terms: in Kantian terms, the claim is that some truths are grounded exclusively in meaning; the positivist claim is that some statements are true by postulation. Quine's response to the positivist challenge appears in "Carnap and Logical Truth" (1954) where he argues that viewing any statement as true by postulation is committing a genetic fallacy.

*(b) The genetic fallacy argument.* The genetic fallacy is, traditionally, a fallacy of relevance: the genesis of a statement *Z* has nothing to do with the rationality of upholding *Z*. Quine's genetic fallacy argument proceeds as follow. Suppose a statement *Z* made it into our corpus of knowledge on grounds of analyticity. What is the significance of this? It shows that at a particular moment in the history of our corpus, *Z* was accepted based on postulation. But this has nothing to do with the reasons for maintaining (or readmitting) *Z* at a later stage in the development of our corpus. Analyticity and syntheticity are no more than genealogical properties of statements—significant *historically*, but not *normatively*. Normatively, our task is to develop a correct theory of the world. Eventually, our postulates must earn their place in our theories not just in virtue of being convenient but also in virtue of being true.[31]

A similar argument by J.G. Maaβ[32] was presented in 1790 against Kant's version of the analyticity claim. Whether a given statement satisfies Kant's conditions on analyticity varies, according to Maaβ, from speaker to speaker and from one occasion of speech to another:

---

[31] The requirement of establishing the truth of postulates is commonplace, according to Quine: any theoretical hypothesis is, at the time it is made, no more than a conventional postulate (being justified by nothing more than "the elegance or convenience...[it] brings to the containing body of laws and data"), and therefore the process whereby we test it for agreement with experience is just the process whereby a conventional postulate is tested for truth—"Carnap and Logical Truth," p. 396. The need to prove the truth of postulates arises also in mathematics, where postulates are often thought of as *obviously* (rather than conventionally) true. Thus, Georg Cantor, for example, held the well-ordering principle to be a self-evident postulate, only later to require a rigorous proof of this principle. See Gregory H. Moore, *Zermelo's Axiom of Choice: Its Origins, Development, and Influence* (New York: Springer, 1982), sections 1.5 and 1.6.

[32] Maaβ's article, "On the Highest Principle of Synthetic Judgments in Relation to the Theory of Mathematical Certainty," appeared in the second volume of the anti-Kantian journal *Philosophisches Magazin*; see Henry E. Allison, *The Kant-Eberhard Controversy* (Baltimore: Johns Hopkins, 1973), pp. 41-45.

some speakers arrive at '*A* is *B*' by being taught that the concept of *B* is included in the concept of *A*, others, by observation; some speakers ***extend*** their knowledge by adding '*A* is *B*' to their corpus, others (or the same speakers on other occasions) do not. Kant (through a review by J.G. Schulze[33]) rejected Maaβ's criticism on the ground that relations between concepts are fixed: we can change the reference of our words, but not the reference of our concepts. Quine, however, denies the "fixity" of concepts, and his own account provides a new theoretical basis for Maaβ's criticism.

(c) *The nonfixity of concepts.* The traditional analytic-synthetic distinction involves the assumption that our concepts are fixed: there is a fixed fact of the matter concerning the content (meaning, reference, composition) of our concepts, hence there are immutable relations between them; theories change, concepts are permanent. Against this view, Quine introduces his principle of the nonfixity of concepts. This principle is closely related to his thesis of the inseparability of language from theory, according to which the meaning of terms is inextricably tied with theories in which they occur: the meaning of 'atom' with theories of matter, the meaning of 'light' with optical theories, and so on.[34] It follows from this thesis that change in theory involves change in the content of our concepts, that is, the content of our concepts is not fixed.

Quine's thesis of the nonfixity of concepts is deepened and expanded by Putnam in "The Analytic and the Synthetic." Putnam establishes the nonfixity of concepts on two grounds: (i) a Wittgensteinian analysis of certain scientific concepts as (in effect) "family resemblance" concepts, and (ii) the observation that sometimes scientific concepts emerge from a radical change in theory changed in content but unchanged in identity. Explanation: Putnam observes that some scientific concepts are interwoven in our theories in such a way that their content is determined by the cluster of laws associated with those theories. (The concept of energy, for example, is embedded in a cluster containing the law of conservation of energy, laws concerning chemical, gravitational, and nuclear energy, laws relating energy to mass and momentum, and so on.) A change in some of these laws results in a change in the content of those concepts, but since those concepts are connected with a host of other laws (which presumably do not all change at once), their identity is preserved across time. Whether and in what way a given law cluster concept will change its content is an open question, according to Putnam: nothing in the concept itself can protect it against change. He restricts his analysis to certain "privileged" concepts of physical science: "energy," "light," "straight line," and so on. But (as Putnam is well aware) there is nothing in his analysis to exclude other concepts from exhibiting the same behavior. Indeed, it follows from the genetic fallacy argument that even concepts originally introduced by convention can become interwoven in our theories in the way Putnam describes. (He claims that at every stage in the development of science some concepts must play the role of "fixed points." This claim granted, it does not follow that the same concepts must play this role at every stage, that is, that there are inherently "fixed" concepts.) The view that all concepts are, in principle, sensitive to change in theory is tantamount to Quine's thesis of the inseparability of language from theory. It follows from this claim that no statement is guaranteed to be true by the "fixed" meaning of its terms, that is, no statement is analytic in the traditional sense.

---

[33] Review of the second volume of the *Philosophisches Magazin*; see Allison, pp. 174-75.

[34] Even in literature our use of terms is affected by theory. Thus, in "A View from Elm Street," Umberto Eco points out that: "The bible obliges its readers to interpret a whale as a big fish; contemporary novels presuppose that whales are to be interpreted as mammals. When Melville seeks to play a *double you* and to tell of a whale that must be considered at the same time as both a fish and a mammal, he spends an entire chapter making his readers eager to overlap two different encyclopedias"—Paulo Leonardi and Marco Santambrogio, eds., *On Quine: New Essays* (New York: Cambridge, 1995), pp. 22-36, here p. 24.

This completes my outline of a defense of NAS: exegetically, Quine is strongly committed to NAS; historically, NAS is a revolutionary thesis of much philosophical interest; and substantively, NAS is supported by three epistemic principles: (i) the principle of maximizing the maneuverability of our cognitive resources, (ii) the principle of the normative insignificance of the genesis of our theories, and (iii), the principle of the nonfixity of concepts.

*B. Second solution: withdraw CP.* The motivation for this solution is straightforward: If NAS is worth preserving and CP is incompatible with NAS, why not eliminate CP?

*Disputation.* We cannot eliminate CP *tout court*, since some positive model of knowledge is needed to validate NAS. Furthermore, CP, in its present form, makes two indispensable contributions to Quine's theory: *(1) it gives structure to his holism, (2) it provides a link between knowledge and reality in his epistemology.*

*(1) Structuring holism.* Without something like the center-periphery duality, Quine's system of knowledge is a formless body of haphazardly interrelated statements. Such an amorphous holism is obviously unacceptable. First, it lacks explanatory power: to give an explanatory account of knowledge is to impose a meaningful structure on its elements. Second, it renders our body of knowledge utterly unmanageable: a structureless object can be grasped either in its entirety, hence in one act of cognition, or not at all; but our system of knowledge is obviously too large, too complex, and too open ended to be grasped in one act. Conclusion: a viable holism must impose a manageable structure on our body of knowledge; that is, a viable holism is not just a structured holism; it is a structured holism with a manageable structure.[*]

CP introduces both structure and simplicity into Quine's holism. It postulates a relatively simple structure of knowledge: a field or a sphere with two distinguished zones. The one zone—*center*—is the center of interconnectedness, the other zone—*periphery*—marks the outer reaches of the system. From the center, our body of knowledge stretches outward toward the periphery, moving gradually from the universal to the particular, from the general to the specific, from the abstract to the concrete, from the conventional to the factual, until reaching its outer boundaries. The periphery is linked to the center by a network of cognitive routes (routes of deductive and statistical inference; routes of abstraction and generalization; routes of application and instantiation; routes of revision, explanation, justification; and the like), and it is through this network of roads that information, evidence, and other epistemic commodities move across the system.

*(2) Anchoring knowledge in reality.* By itself, Quine's holistic conception of knowledge with its principles of interconnectedness, universal revisability, NAS, pragmatism, and so on is compatible with a coherentist or even an idealistic epistemology. Any system of statements will satisfy these principles, provided certain conditions having to do with the internal structure of the system are satisfied. The transformation of Quine's holistic theory of knowledge into a theory

---

[*] Dummett's criticism of NAS in *Frege: Philosophy of Language* is based on the (mistaken) assumption that NAS itself implies a structureless holism. Some of Jerry Fodor and Ernest Lepore's criticisms in *Holism: A Shopper's Guide* (Cambridge: Blackwell, 1992) are also directed at a structureless holism.

of knowledge of reality is accomplished by CP. CP delineates the inner structure of our system of knowledge in relation to reality. It anchors our corpus of knowledge in reality, on the one hand, and in language (the mental), on the other. Our corpus is linked to reality through the periphery, and the impact of reality permeates the corpus due to interconnections between its elements. These interconnections are regulated in the center, where the influence of language, method, mind, pragmatic norms, and so on is most strongly felt. Changes may take place anywhere in the sphere, but all changes must square with reality (directly or indirectly) through the periphery. Even a theory as highly removed from experience as Einstein's general relativity theory must square with experience through the periphery, that is, via crucial experiment. If we call the requirement that our system of knowledge be anchored in reality *realism* (or *minimal realism*), then a theory of knowledge based on NAS requires something like CP to render it (minimally) realistic.

In making these contributions to Quine's epistemology, however, CP assumes the form of an ultratraditional model: one fixed center—the site of logic, mathematics, and philosophy—surrounded by scientific theories whose supporting evidence (in the form of experiential reports) is transported from the periphery. We may liken this structure to a traditional industrial economy: one metropolis—the locus of capital, services, means of communication—surrounded by industrial sites whose raw materials are transported from quarries and mines across the country. The system as a whole is interconnected, but the norms governing its center and periphery are radically different: the norms governing statements in the center have to do with the quality of "general services," those governing statements in the periphery—with the quality of "raw materials" (the embeddedness of the system in reality). It is this rigid division of the cognitive norms that is irreconcilable with NAS.

## IV THIRD SOLUTION FROM AN ABSOLUTIST-STATIC MODEL TO A CONTEXTUAL-DYNAMIC MODEL

Our analysis suggests that NAS is an innovative thesis of much philosophical interest, while CP makes an important contribution to demonstrating its epistemic viability. Neither thesis should be eliminated, yet the two stand in a fundamental tension. A natural way of resolving the problem is by reconfiguring CP, that is, identifying those elements of CP which are responsible for the conflict and replacing them by other elements in agreement with NAS.

My solution is based on the following observation: there is no conflict between the idea, systematized by CP, that the two main sources

CALCUTTA UNIVERSITY LIBRARY
ALIPUR
CAMPUS

of knowledge are language and the world, and the idea, central to NAS, that the division between them within a given statement is neither fixed nor lasting. Rather, the root of the conflict is the assumption, implicit in CP, that each statement, and, by extension, each theory, is located within a fixed region of the sphere—within a fixed distance from its center and boundary—and this distance is determined by the ratio of conventional to experiential factors in its truth and justification. The presumption of a fixed location involves the requirement that a fixed range of experiences and a fixed range of conventions be associated with each statement, but exactly this is contested by NAS. It is thus not the existence of two distinct sources of knowledge, metaphorically referred to as "center" and "periphery," that is the source of the conflict; rather, it is the postulation of fixed factual conditions and fixed conventional conditions for the truth and justification of each statement that is responsible for it. The fixity requirement has two dimensions: (i) at any given point in the duration of our corpus each statement occupies a fixed (unique, absolute) position within the sphere, and (ii) while in the course of history a statement may change its exact location (for example, as a result of a re-axiomatization of a theory to which it belongs), its general area as well as its rough distance from the center and the periphery remain the same. This is true especially of observation statements, which always occupy the periphery, and of logico-mathematical statements, which are always in the center.

My solution to the inner conflict in Quine's theory lies in replacing this traditional, static, and absolutist model by a new, dynamic, and contextual model. There are still two zones, corresponding to the two basic constituents of human knowledge, but no fixed distribution of statements and theories within the model. As a theory of the formal laws governing structures of objects—for example, the laws relating identity to cardinality—logic is in the periphery; as a theory of the logical laws governing statements and sets of statements, logic is in the center. As a theory saying that a structure of all identical objects is a structure of one object, logic is about the world; as a theory saying that the statement 'All objects are identical'[38] implies the statement 'There is exactly one object', logic is about language. As a theory of the formal structure of "real" objects (configurations of objects), logic is anchored in reality; as a theory of the formal structure of our thought of objects, logic is embedded in the mental. A similar duality exists in other disciplines. Take zoology, for example: as a theory of the hierarchy of life, it is in the periphery; as a system of concepts of life (as a hierarchy of zoological concepts), it is in the center.

---

[38] Or, rather: 'All objects are identical to some object'

Normatively, the project of knowledge consists of two tasks: the task of constructing a serviceable set of conceptual tools, and the task of constructing a correct and informative theory of the world. Accordingly, our system of knowledge is measured by standards of two types: substantive standards—truth, confirmation, prediction, explanation; and pragmatic standards—economy, convenience, utility. In constructing a system of knowledge, we move back and forth between the center and the periphery: we set out to improve the overall usefulness of our conceptual apparatus, and we set out to verify the correctness of our theories. Each theory in our corpus must square off with some facet of reality: logic with the formal behavior of objects, physics with their physical behavior. Given a conflict with reality, we are given great latitude in devising a solution. We may resolve the conflict by making changes in our physical theory of the world, keeping logic "fixed" in the background; or we may resolve it by changing our theory of the formal structure of the world, keeping the physical laws "fixed" in the background. We may resolve the conflict by placing physics in the "front" (periphery), logic in the "rear" (center); or we may resolve it by placing logic in the front, physics in the center. By holding our notion of physical object fixed, we constrain the variability of our logic; by holding the laws of logic fixed, we restrict the variability of our physics. Ontology plays a similar role, as do other branches of philosophy. Philosophy, on this picture, is not above (or below) the sciences; rather, philosophy—through its theories of logic, object, truth, rationality, and the like—faces the court of knowledge both in the center (on the level of concepts and method) and in the periphery (on the level of fact), just like other sciences. And like other sciences, philosophy has its own identity, methods of inquiry, and relative autonomy. (Philosophy is not reducible to empirical psychology.)

While Quine's narrow naturalism is rejected by our model, many of his early theses are integrated into it: NAS, the interconnectedness of knowledge, the universal revisability of beliefs, the inseparability of language from theory, (minimal) realism, scientific empiricism, antifoundationalism, and certain important features of the center-periphery duality. In particular, the new model continues to make the two main contributions of the original CP: (i) structuring the interconnections between statements (and between theories), and (ii) linking our system of knowledge to reality. Unlike the old model, however, the new model is dynamic and contextual: at each stage in the development of our corpus, each science lies closer to the periphery in some contexts, closer to the center in others, and in the course of history, sometimes one constellation of theories, sometimes another, occupies a more prominent place in the periphery.

Methodologically, the new model is both (highly) pragmatist and (moderately) metaphysical: *pragmatist* in the idea that given a conflict, we have great latitude in maneuvering our cognitive resources; *metaphysical* in the idea that some conflicts with reality are best met by placing logic, philosophy, and mathematics in the periphery. Such a combination of metaphysics and pragmatism is, according to Quine, characteristic of NAS: "One effect of abandoning [the traditional dogmas] is...a blurring of the supposed boundary between...metaphysics and natural science. Another effect is a shift toward pragmatism."[37]

## V THE NEW CP: AN OUTLINE AND CRITICAL REMARKS

To clarify the structure, scope, and ramifications of the new CP, both as a critical interpretation of Quine's theory and as a model standing on its own, I shall offer a concise outline of the model. The outline will center on the two dimensions along which concepts, statements, and theories shift their position within the model: the *time dimension* and the *context dimension*. It is quite clear that the addition of these dimensions to the old CP does not detract from its ability to represent the Quinean themes of interconnectedness, universal revisability, inseparability of language from theory, pragmatism, and so on. The outline will focus, therefore, on the movement of statements and theories along the two dimensions, the sense in which logic and philosophy lie (in certain respects and at certain times) in the periphery, the normative principles (associated with NAS) underlying the model, and some of the model's ramifications for outstanding philosophical problems. In drawing the outline, I am aiming at clarity and, to a certain extent, systematicity, but not at a comprehensive or exhaustive account. While the outline expands the exegetical basis of my solution by drawing upon Quinean ideas not considered in earlier sections, it is important to remember that the model is intended as a *critical* interpretation (and development) of Quine's theory, and as such it diverges from Quine on certain significant points.

---

[37] "Two Dogmas of Empiricism," p. 20.

The center-periphery metaphor is interpreted as a "family" of dualities: "fact versus convention," "objects versus concepts," "front versus rear of knowledge," and so forth.

Two dimensions of change: the most distinctive characteristic of the new CP (compared with the old one) is the "movement" of cognitive elements along two axes: *time* and *context*. Diachronically, the new model is dynamic, that is, as our body of knowledge develops, the deployment of concepts, statements, and theories within it changes in response to changing circumstances. Synchronically, the new model is contextual, that is, at each stage in the development of our system each statement, concept, and theory falls within a multiplicity of contexts, and viewed from the varied perspectives of these contexts, its position within the system changes. In the extreme case, statements (concepts, theories) shift their position from the center to the periphery and from the periphery to the center.

A. *Contextual change.* We can distinguish two (complementary) ingredients of contextual change: (*1*) *multiplicity of perspectives*, (*2*) *multiplicity of factors.*

*(1) Multiplicity of perspectives.* Change in perspective may lead to change in orientation: viewed from one perspective, a concept, statement, or theory is world oriented, viewed from another, language or mind oriented; viewed from one perspective, a given subject matter is factual, viewed from another, mental or conceptual. I shall consider two sources of such change in perspective: *(a) the two-sidedness of language, (b) the plurality of human interests.*

*(a) The two faces of language; semantic ascent and objectual descent.* One source of the multiplicity of perspectives is the two-sidedness of language. It is an old observation (made by medieval philosophers) that sentences in general can be read both in the linguistic mode and in the factual mode. Take, for example, the sentence 'Whales are mammals'. Read in the linguistic mode, this sentence says that 'whale' is subsumed under 'mammal'; read in the factual mode, it says that whales possess the characteristic properties of mammals. Read in one mode this sentence is language oriented, read in the other, world oriented. One important vehicle of such change in orientation is the truth predicate. The truth predicate enables us to speak about the world by speaking about language, and to speak (in certain ways) about language by speaking about the world. Instead of saying that snow is white, we can say that 'Snow is white' is true, and instead of saying that 'Snow is white' is true, we can say that snow is white. Quine introduced the term *semantic ascent* for one di-

rection of this shift; I shall use *objectual descent* for the other direction.

*(i) Truth is both language and world oriented.* The dual principle of semantic ascent—objectual descent is encapsulated in the truth schema. But the truth schema itself, according to Alfred Tarski and Quine, allows both a linguistic and a factual reading. Read in the linguistic mode, the truth schema is a disquotational schema; read in the factual mode, a correspondence schema. Under one reading, truth has to do with an intralinguistic relation; under the other, with a relation between language and the world. While the two readings yield the same *T* sentences, their explanation of truth is radically different. The disquotational account of truth is linguistic, the correspondence account, metaphysical; the disquotational notion of truth is language oriented, the metaphysical notion, both language and world oriented.

*(ii) Objectual descent: logic is world oriented.* The principle of objectual descent provides one explanation of logic's presence in the periphery. It follows from this principle that the logical laws, despite their appearance (for example, in the case of sentential logic) as laws governing the behavior of language, reflect certain regularities in the behavior of the world. In Quine's words: "Logical theory, despite its heavy dependence on talk of language, is...world-oriented rather than language-oriented; and the truth predicate makes it so.... [T]he truth predicate is...doing an active job of separating logic from language."[illegible] And elsewhere he says: "Consider...the logical truth 'Everything is self-identical', or '$(x)(x=x)$'. We *can* say that it depends for its truth on traits of the language (specifically on the usage of '='), and not on traits of its subject matter; but we can also say, alternatively, that it depends on...[a] trait, viz., self-identity, of its subject matter, viz., everything."[illegible] On the second interpretation, the law of self-identity is world oriented.

*(iii) The epistemic priority of the factual: the immanence of knowledge.* While statements and theories within the model are both fact and language oriented, epistemically their factual orientation takes priority over their linguistic orientation. The epistemic primacy of the factual is tantamount to Quine's thesis of the *immanence of knowledge.* This thesis says that we know the world through our theories—that is, from a standpoint internal to our theories—but speaking from

---

[illegible] *Philosophy of Logic*, p. 97.

[illegible] "Carnap and Logical Truth," p. 390.

within a theory is speaking factually. We use words (in our theories) to speak about objects, and in so doing we give priority to their factual reading. (We can view this as the principle of "use over mention"). The epistemic priority of the factual over the linguistic means that all knowledge is, in a certain basic sense, fact or periphery oriented. This principle is reflected in Quine's claims that logic is "world-oriented rather than language-oriented," and that mathematics and ontology are about "what there is."

*(iv) Philosophy is fact oriented.* The immanence of knowledge means that all sciences are fact oriented. We can explain the way this principle applies to philosophy (even beyond logic and ontology) by contrasting the factuality of philosophy with its alleged conventionality. Consider Kant's epistemology once again. The motivation for his "Copernican revolution" is, as we saw in section II, largely pragmatic: all attempts to establish knowledge based on the traditional conception of the relationship between the knower and the world have failed; let us see whether by turning this relationship around we shall not be able to arrive at a better theory. But the theory constructed based on this move makes *factual* claims about its subject matter, namely, human knowledge. Speaking in terms of aims of theory, we can say that Kant's theory aims not just at an expedient account of human knowledge, but also (and indeed primarily) at a correct account. Kant's goal is to identify the real conditions for the possibility of knowledge (not some convenient pseudo conditions), to give a correct analysis of the basic structure of our cognition (not a convenient fiction about its structure), and to provide a valid proof of the possibility of human knowledge (not the semblance of a valid proof, that is, a chain of statements that can conveniently be imagined to be a valid proof). Kant aims at a correct theory of the structure and possibility of human knowledge, and in this sense his theory, and philosophy more generally, is oriented toward the periphery. This is perhaps the deepest sense in which philosophy lies in the periphery. Philosophy seeks to obtain an understanding of knowledge, truth, morality, valid inference, and so on as they in fact are: their true factors, true structure, true possibility, true difficulties—all as they truly or factually are, not as we may conveniently wish, or imagine, or postulate them to be.

*Critical note*: the factuality of philosophy means not that philosophy is less pragmatic than other sciences, but rather that the norms of truth (accuracy, absence of error, and so on) are as central to philosophy as they are to other sciences.

(*b*) *Multiplicity of interests.* Another important source of the multiplicity of perspectives is the multiplicity of interests. The multiplicity of interests leads to the development of multiple theories of the same subject matter. Take the subject of correct reasoning, for example. By studying correct reasoning from a purely theoretical standpoint, we arrive at a "pure" logic; while by studying it from a point of view encompassing both the practical and the theoretical, we arrive at a (general) theory of rationality. Differences in perspective may lead to radical differences in our conception of a given subject matter: by thinking of the logical laws as laws of the interaction of physical properties (properties of physical states), quantum logicians have come to view them as empirical; by thinking of these laws as governing the manipulation of symbols, the formalists came to view them as conventional. Different perspectives may either agree or disagree with one another: the intuitionistic viewpoint on logic competes with the classical viewpoint, whereas the modal viewpoint leads to its extension. Different theories of the same subject matter are located in different areas of the field: a psychological theory of reasoning lies closer to the periphery (in the traditional sense of site of experiment and observation), a "pure" logical theory closer to the center.

*Critical note:* not every point of view on every subject matter is validated by the model. That is, the model does not represent a methodology of "anything goes." On the contrary: it is a central principle of the new model (closely related to its representation of NAS) that every theory is subject both to the norms of truth (accuracy, avoidance of error, justification, explanation) and to the norms of efficiency (simplicity, unity, generality, fruitfulness, and so on). And this means that the standards for the acceptance of a given viewpoint as a basis for a new theory are higher in the new model than in either the traditional or the positivist models. We can express this by saying that each theory is subject not just to the norms of utility but also to the norms of truth, and not just to the norms of truth but also to the norms of utility. As a result, most points of view on a given subject matter are rejected by the model (or rather by a world theory concordant with it) either as conducive to error or as inexpedient, unfruitful, and so on. Nevertheless, the principle of a multiplicity of viewpoints is affirmed by the model.

(2) *Multiplicity of factors.* The contextual mobility and malleability of subject matters, statements, and theories within the model is not, primarily, a subjective matter, having to do with our personal choices, whims, and proclivities, but (to a large extent) an objective

principle, having to do with the multifariousness of the world and the plurality of factors constraining the knower.

*(a) Multiplicity of factors constraining the knower; knowledge is not a mirror representation of reality.* The view that knowledge is largely shaped by factors having to do with the knower is characteristic of many epistemologies. Kant, and, in a different way, contemporary psychologists, linguists, and cognitive scientists, view mental structures as playing a crucial role in determining our theory of the world. The logical positivists emphasize the importance of pragmatic factors in shaping our knowledge. Following Thomas Kuhn, many philosophers, historians, and sociologists of science see scientific knowledge as largely determined by social, psychological, political, and environmental factors. The new model accepts the general principle underlying these epistemologies. It follows from this principle that knowledge is not a simple picture of reality (words are not mere labels of objects; sentences and theories are not mirror images of "facts"), but numerous factors residing in the knower intervene between our world theory and the world.

*Critical note:* the existence of factors residing in the knower does not conflict with the applicability of the norms of truth. A street map is not a mirror image of a city, but relative to certain well-defined goals and standards, it is either a correct or an incorrect image of it. Likewise knowledge, in the present model, is not a *mirror* representation of reality, but it is (or is required to be) a *correct* representation of it. This inner duality is inherent in NAS: our statements and theories, according to NAS, are both conventional and factual; both governed by the norms of "simplicity, familiarity, scope, and fecundity,"[40] and subject to the norms of truth and evidence. This duality extends to ontology: the objects of our theories, according to Quine, are both "posits" and "what there is"; as "posits" their knowledge conforms to the norms of utility, as "what there is," to the norms of truth. The model explains this two-sidedness of knowledge by its double origin in the world and the mind. The interaction of "worldly" and "human" factors is not treated as a mystery by the model, but as something for theories within it (epistemology, psychology, biology, sociology, and so on) to account for.

*(b) Multiplicity of factors concerning the world; the factuality of mathematics, logic, and ontology.* The model admits not only multiple factors re-

---

[40] See "Posits and Reality," in *The Ways of Paradox and Other Essays*, pp. 246-54, here p. 247.

siding in the knower, but also multiple factors residing in the world. The multiplicity of worldly factors is related to the diversity of the world. Different factors—connected with different parts, features, and aspects of the world—contribute to our knowledge, and the plurality of these factors partly explains the plurality and diversity of our theories. It is commonplace to say that physics investigates a different "region" of reality from biology; but in the new model this view extends to logic, mathematics, and philosophy. Scientists, from Copernicus to Einstein, regarded physical reality as exhibiting not only material features but also formal (mathematical) structure. The factuality and distinctiveness of mathematics are predicated, in part, on the reality and distinctiveness of such structures. The view that logic is grounded in special features of reality is suggested by Quine when he says: "admittedly [logical truth] depends upon none of those features of the world that are reflected in the lexical distinctions; but may it not depend on other features of the world, features that our language reflects in its grammatical constructions rather than its lexicon?"[41] Ontology, according to Quine, is also concerned with certain features of reality, namely, properties of objects of an especially broad nature: "The question what there is is a shared concern of philosophy and most other non-fiction genres.... What distinguishes between the ontological philosopher's concern and [the zoologist's, physicist's and mathematician's concerns] is... breadth of categories."[42]

*Critical note*: in considering the claim that logic and mathematics are (in certain respects) located in the periphery, it is important to note: (i) this claim does not imply that logic and mathematics are empirical; (ii) this claim is not tantamount to Platonism. In fact, neither extreme empiricism (the view that apriori, all knowledge is restricted, or reducible, to empirical knowledge) nor Platonism (the belief in a "third realm") is compatible with the model.[43]

*(c) Multiplicity of factors underlying institutions; reconciling holism and compositionality.* The principle of multiplicity of factors applies not just to the knower and the world but also to the institutions of knowledge. Take the institution of language, for example. The model

---

[41] *Philosophy of Logic*, p. 95.

[42] *Word and Object*, p. 275.

[43] Quine himself emphasizes that empiricism is not "hard wired" into his model (see "Two Dogmas," p. 44). His espousal of scientific empiricism is motivated by pragmatic considerations, but these considerations do not apply to mathematics. This explains why, for Quine, mathematics is nonempirical, albeit continuous with empirical science.

permits us to affirm both the holistic nature of language (the fact that the meaning of any term may, in certain respects and under certain conditions, be affected by that of any other term) and its compositionality (the fact that the meaning of complex terms depends, in some systematic way, on the meaning and mode of composition of simpler terms).[44] The two are allowed to co-exist as two independent factors whose complex pattern of interaction is a matter for linguistics, psychology, sociology, the philosophy of language, and other theories within the model to investigate. By combining holism with discrete constituents, the model is "organic" in Dummett's sense.

*(d) Observation: multiplicity of factors as richness of structure; enhanced explanatory power.* The multiplicity of factors characteristic of the new model amounts to richness of structure. The model portrays the process of knowledge as an intricate process, involving numerous factors and constituents, dynamically interacting with each other. By allowing rich structures, the model's ability to deal with complexity and diversity is enhanced. In particular, the model is capable of reconciling elements which, in a simpler model, would appear to conflict with each other—for example, holism and compositionality, as discussed above. Indeed, the difference between the old model and the new one can be explained by richness (poverty) of structure. By taking into account a relatively small number of factors, the old model delineates a very simple structure of knowledge: a static and rigid structure with a "fixed" center and a "fixed" periphery, simulating the analytic-synthetic division. By adding two new dimensions (two new clusters of factors) to the old model, the new model adds movement and complexity to its structure, relativizing its center-periphery duality to context and time and reconciling it with NAS.

*B. Change in Time.* I shall note three types of diachronic change, corresponding to three aspects of the center-periphery duality: *(1) factual orientation versus conceptual orientation, (2) "front" versus "rear" of knowledge, (3) variant versus invariant ("fixed") constituents of knowledge.*

*(1) Change from factual to conceptual orientation (and vice versa). (a) Developmental change.* At each stage in its development, each (viable) science has both factual and conceptual interests, but at different periods in its development different interests may play a more, or less, dominant role. How to classify each period in the history of a given

---

[44] My use of 'compositionality' here is broader than usual. It covers any semantics which analyzes complex structures by a finite algorithm based on their parts, including, in addition to Tarskian semantics, Leon Henkin's functional semantics for partially-ordered quantifications, Hintikka's game-theoretic semantics for IF languages, and others.

science is a matter for the historian and philosopher of science to determine, but during periods of major conceptual advances the science moves closer to the center, during periods of intensified factual investigations (either in the sense of experiment and observation, or in the sense of theoretical development) it veers toward the periphery. Logic and mathematics, too, alternate between the factual and the conceptual: the introduction of the modern quantifiers by Gottlob Frege, for example, is naturally viewed as a conceptual contribution, while Kurt Gödel's discovery of the incompleteness of arithmetic is, in an important sense, a factual contribution. All scientific revolutions, according to Kuhn, involve major changes in conception and methodology (both practical methodology and theoretical methodology); such changes are represented in the model as occurring (largely or partly) in the center.

*(b) Change in grounds and significance.* The same statement or theory can be accepted based on one ground during one period, based on another during another (the genetic fallacy argument); similarly, its significance may vary in the course of time. For example, it is quite natural to view non-Euclidean geometry as moving from a phase (in the nineteenth century) in which its significance was primarily conceptual (independence of the parallel postulate and generalization of Euclidean geometry to Reimannian geometry) to a phase (in the twentieth century) in which its role and significance are largely factual (a theory of the curvature of physical space).

*(2) Movement from "rear" to "front" of knowledge (and vice versa). (a) Movement of whole disciplines.* In the battle against nature sometimes one discipline, sometimes another, occupies the front. It is common to think of the natural and social sciences as occupying this position: physics in 1687 and 1905, biology in 1859, psychology in 1900, and so on. Traditionally, logic and philosophy are regarded as permanently located in the center, "setting the stage" for the battle against nature rather than taking an active part in it. The NAS methodology represented by the model challenges this tradition. Speaking of logic's role in contemporary physics, David Finkelstein says: "[Logic is a] dynamic ingredient in the physical theory, *an actor rather than part of the stage.*"[8] In the new model, logic plays both the role of an actor and that of managing the stage. As an actor, the challenges taken up

---

[8] "Matter, Space and Logic," p. 199.

by logic are of a general and fundamental nature: the challenge of valid inference (transmission of truth from sentences to sentences), the challenge of bivalent and nonbivalent phenomena, and the like. Philosophy, too, is assigned a double role. One of the challenges posed by nature to philosophy is the skeptical challenge, which questions humanity's ability to break out of its "mental cage" and attain "true" knowledge of the world. Philosophy took up the skeptical challenge in René Descartes's *Discourse on Method* and *Meditations*, Kant's *Critique of Pure Reason*, and elsewhere. Quine's 1968 manifesto, "Epistemology Naturalized,"[*] calls for fighting the skeptical battle on the psychological rather than the philosophical front. Critics of Quine's naturalism claim the challenge cannot be adequately met by psychology. The new model takes no apriori stand in this controversy. The choice of strategy is an open question, to be decided based both on the nature of the challenge and on the availability of resources. The balance of these factors may vary from era to era, and with it the choice of strategy may vary.

*(b) Movement of special issues.* Not only whole disciplines but also single questions can shift their position from the rear to the front. Take the *Entscheidungsproblem*, for example. As posed by David Hilbert in 1900, this problem concerned the existence of a finite procedure for determining the solvability of Diophantine equations. But though a problem of considerable interest for mathematics, it was far from being one of the central issues facing our system of knowledge at the turn of the century. In the 1930s and 1940s, however, the *Entscheidungsproblem* became the focus of a serious challenge (or group of challenges) to our system. This challenge—subsumed under the headings of completeness, decidability, computability, recursiveness, and so on—concerns the scope and limits of a broad class of procedures and functions of special importance for knowledge. With the work of Gödel, Alonzo Church, Alan Turing, Emil Post, John von Neumann and others, the *Entscheidungsproblem* moved to the front of knowledge, leading to some remarkable gains in the battle against nature, as well as to some remarkable losses.

*(3) Change from being a "variant" to being an "invariant" ("fixed") constituent of knowledge (and vice versa).* The idea of the center and the periphery as representing the "fixed" and "varying" constituents of knowledge is a traditional idea: our language, logic, and mathemat-

---

[*] *Ontological Relativity and Other Essays* (New York· Columbia, 1969), pp. 69-90.

ics are fixed, experiential knowledge is variable and changing. The new model relativizes this duality both to context and to time. The underlying principle is Otto Neurath's: we build our system from within, by shifting our foothold from one area within it to another. Placing our foothold in area *A*—that is, holding the concepts, laws, and theories of area *A* fixed—we rebuild area *B*; shifting our foothold to area *B* (or to an area intersecting with *B*), we change area *C*; and so on.[47] The area we hold fixed at time interval *t* is the center at *t*, the area we change at *t*, the periphery at *t*. This Archimedean principle is both dynamic and contextual: we can use *A* as a lever for changing *B* both at a particular time and in a particular context. Among the many ramifications of this principle are:

(*i*) *Explanation of continuity through change.* One of the main challenges facing any epistemology is the explanation of continuity through change. The problem is especially difficult for theories permitting radical change, like Quine's theory with its universal revisability and inseparability theses, and Kuhn's theory of scientific revolutions as paradigm shifts. In the current literature, the problem is variably referred to as the "incommensurability" problem, the "meaning variance" problem, the "change of theory is change of subject" problem, and the like. The new model is well equipped for handling this problem. Think of the "fixed" area of our system at *t*—the center at *t*—as containing those elements which are stable at *t*. The center is what "glues" the system together, the periphery is the locus of change and revision. During periods of "normal" development, logic, methodology, and general scientific principles are held fixed in the center; during periods of radical change, it is conceivable that some combination of simple observation statements, everyday concepts, common-sense maxims, and logical and theoretical elements (not undergoing change) constitute the center. The exact dynamics of variance and invariance is a research project for various disciplines within the model, but the model itself offers a flexible framework for the explanation of stability through change.[48]

---

[47] While the question of circularity arises in this situation, the applicable circularity criterion will have to take into account the nonfoundational nature of the model.

[48] A similar approach to the stability problem is suggested by Mary Hesse in "Duhem, Quine and a New Empiricism," in Sandra G. Harding, *Can Theories Be Refuted? Essays on the Duhem-Quine Thesis* (Boston: Reidel, 1976), pp. 184-204, and by Gochet, chapter 1, section 6.

*(ii) Foundations without foundationalism.*[49] By allowing all sciences to occupy the same "structural" positions within the "sphere" (that is, center, periphery, and intermediate positions at varying distances from the two), the model subjects all sciences to the same (highly general) norms. Among these are the norms of *critical evaluation* and *substantive justification.* If we regard these norms as the *foundation norms,* then every science within the model—including sciences that in traditional models are placed at the bottom of the foundational pyramid—requires a foundation. Neurath's principle allows the model to provide such a foundation. Each science stands in a multitude of normative, conceptual, practical, and theoretical relationships to other sciences, and by selecting an appropriate standpoint among these sciences, we can use their resources to provide it with a foundation. While an apriori, universal foundation—that is, a foundation guaranteed in advance for all contexts and all times—is at variance with the model, its norms of unity, generality, and economy favor broad foundations over narrow foundations.

*(iii) A standpoint for philosophy ("immanent transcendence").* The view that philosophy requires an external standpoint has led philosophers like Ludwig Wittgenstein and Richard Rorty to deny the possibility of a substantive philosophy. A substantive theory of knowledge, for example, requires a viewpoint encompassing our world theory in its entirety, but such a viewpoint is not available to humanity. The new model allows us to reconcile the immanence of knowledge with the transcendence of philosophy. It suggests the possibility of a very broad standpoint: an area of knowledge from which we can view all other areas, both in their relation to each other and in their relation to the world (as seen from the chosen area). The chosen area may offer an indirect view of some of its own regions, as does arithmetic with respect to its own syntax and certain mathematical and scientific theories with regard to their own ontology.[50] Thus, although an absolute transcendent viewpoint is ruled out by the model, a relative transcendent viewpoint—a point of view transcending all but a makeshift center—is compatible with it. Such a point of view is, indeed, characteristic of the model itself. The model offers a critical outlook on our system of knowledge from a standpoint within it, namely, that of a late twentieth-century view of

---

[49] This phrase is taken from the title of Stewart Shapiro's *Foundations without Foundationalism: A Case for Second-order Logic* (New York: Oxford, 1991).

[50] See Quine, "Ontological Relativity," in *Ontological Relativity and Other Essays,* pp. 26-68.

Quine's early theses and the conflict arising from an attempt to represent them by a static, absolutist model.

This concludes my outline of the new CP. The model resolves the inner conflict in Quine's theory by upholding the principle of a center-periphery structure while denying the assumption of a "fixed" distribution of statements and theories within this structure. Every statement and theory lies in multiple areas of the structure, and as such it is not restricted to the norms of any one area. In particular, philosophy is subject both to the norms of simplicity, utility, convenience, and economy (the norms associated with the center), and to the norms of truth, evidence, justification, explanation (the norms associated with the periphery and intermediate sections). As a result, philosophy is both conventional and substantive: both a practical discipline in the positivist tradition, and a factual, theoretical discipline in the image of metaphysics and science.

GILA SHER

University of California/San Diego

# CAN A TURING MACHINE KNOW THAT THE GÖDEL SENTENCE IS TRUE?[*]

A Turing machine can *know* what it can *prove*, that is, deduce from axioms using well-defined rules of inference. Its axiomatic database may be large, and its rules of inference may be efficient and powerful. But for a Turing machine, *knowability* = *provability*. When truth and provability part company, as in the case of Gödel's incompleteness theorem, a machine can cope with one but not with the other.

The short answer to the question in the title is "no," because if *Peano arithmetic* (PA) should turn out to be inconsistent, the Gödel sentence *G* would be false not true. The truth value of *G* depends on an unknown: the consistency of PA. In any case, truth and provability diverge: for if PA is consistent, *G* is true and unprovable; and if inconsistent, provable and false. On the other hand, conditionalizing on PA's consistency yields 'if...then' sentences that are unconditionally true or false, and here the connection between truth and provability is more subtle. In section II, two conditional sentences are examined:

A1: If PA is consistent, then *G* is not provable.
A2: If PA is consistent, then ~*G* is not provable.

Both sentences are true, the difference being that the formal version of A1 in the language of PA is a theorem, whereas the formal version of A2 to the best of our knowledge is not. Although a "PA Turing machine," programed to run through and check proofs in PA, will be able to enumerate all PA theorems, and hence will *know* A1 in the sense that it can prove it, the machine, unlike a human being, will be incapable of recognizing, grasping, or otherwise perceiving the truth of A2. The concept of "unprovable but true" is beyond it. Or so I shall argue.

There have been two recent attempts, by John Lucas and Roger Penrose,[1] to use Gödel's theorem to show that any computational device is limited, compared to human beings, in what it can know or understand. Their arguments, which if correct would imply that the program of strong AI was ultimately not viable, have provoked vigor-

---

[*] The author wishes to thank Anil Gupta, Allen Hazen, Zlatan Damnjanovic, and Alasdair Urquhart for helpful comments.

[1] Lucas, "Minds, Machines and Gödel," *Philosophy*, xxxvi (1961). 112-27; Penrose, *The Emperor's New Mind* (New York: Oxford, 1989), *Shadows of the Mind* (New York: Oxford, 1994).

0022-362X/99/9610/525–32

ous rebuttals from other philosophers.[2] I shall argue along different lines. I focus on the sharp distinction between truth and provability and show that whether PA is consistent or inconsistent, the set of provable sentences differs from the true sentences. I then take up an example of a sentence in PA which is true, but which cannot be categorized as true by a PA Turing machine. The same holds for Turing machines with richer axiomatic bases; for each of them there will be at least one proposition which a human can know to be true, but which the machine lacks the capacity to recognize. At every level, there will be specific cases in which human mental powers outstrip the capabilities of Turing machines.

## I. GÖDEL'S THEOREM

Among the things that Gödel showed is that, if PA is consistent, the concept of proof is representable in PA. This means that, if *Pr* is the set of ordered pairs $<x,y>$, where $x$ is the Gödel number (GN) of the proof of the well-formed formula with GN $y$, then there is an open PA formula:

$$\underline{\quad}x\underline{\quad}y\underline{\quad}$$

which is such that for all $x$ and $y$, $<x,y> \in Pr$ if and only if:

$$\vdash \underline{\quad}\bar{x}\underline{\quad}\bar{y}\underline{\quad}$$

where $\bar{x}$ is $x$ in Peano notation, that is, $0'''''\quad'$.

This is the definition of *weak* representability. So-called *strong* representability[3] means that, for any $x$ and $y$, if $<x,y> \in Pr$ then $\vdash\underline{\quad}\bar{x}\underline{\quad}\bar{y}\underline{\quad}$, and if $<x,y> \notin Pr$ then $\vdash \sim(\underline{\quad}\bar{x}\underline{\quad}\bar{y}\underline{\quad})$. Recursively enumerable sets are weakly representable; decidable sets are strongly representable. Since it is a routine clerical matter to determine which pairs $<x,y>$ belong to *Pr*, *Pr* is a decidable set and is strongly represented by $\underline{\quad}x\underline{\quad}y\underline{\quad}$. On the other hand, although the set *Th* of Gödel numbers of theorems of PA is enumerable because proofs are enumerable, *Th* is not decidable because the nontheorems of PA are not enumerable. The open formula that weakly represents *Th* is $(\exists x)(\underline{\quad}x\underline{\quad}y\underline{\quad})$.

---

[2] See Paul Benacerraf, "God, the Devil, and Gödel," *The Monist*, LI (1967): 9-32; David Lewis, "Lucas against Mechanism," *Philosophy*, XLIV (1969): 231-33, and "Lucas against Mechanism II," *Canadian Journal of Philosophy*, IX (1979): 373-76; G.L. Bowie, "Lucas' Number Is Finally Up," *Journal of Philosophical Logic*, XI (1982)· 279-85; Hilary Putnam, review of Penrose's *Shadows of the Mind*, *The New York Times Book Review* (November 20, 1994), p. 7; also numerous commentators in *Behavioral and Brain Sciences*, XIII (1990): 655-92.

[3] See Moshe Machover, *Set Theory, Logic and Their Limitations* (New York: Cambridge, 1996), p. 223.

Gödel also noted that the set Sub of triples $<w,y,z>$ is decidable, where $w$ is the GN of the formula which results from substituting the number $y$ (in Peano notation) for all instances of free variables in the formula with GN $z$. Furthermore, $w$ is a function of $y$ and $z$, so that the open formula that represents Sub can be written in the form $w = \mathrm{sub}(y,z)$. Construct the formula:

A3: $\sim(\exists x)(\underline{\quad\;}x\underline{\quad\;}\mathrm{sub}(y,z)\underline{\quad\;})$

and let the GN of A3 be $g$. Then the Gödel formula $G$ is:

$G$: $\sim(\exists x)(\underline{\quad\;}x\underline{\quad\;}\mathrm{sub}(\bar{g},\bar{g})\underline{\quad\;})$

Since $\mathrm{sub}(g,g)$ is the GN of $G$, $G$ says of itself that it is not a theorem of PA.

Is $G$ provable in PA? Suppose, for reductio, that it were, and let $h$ be the GN of its proof. Then $<h,\mathrm{sub}(g,g)> \in Pr$. By representability:

A4: $\vdash \underline{\quad\;}\bar{h}\underline{\quad\;}\mathrm{sub}(\bar{g},\bar{g})\underline{\quad\;}$

But from $\vdash G$ we derive:

A5: $\vdash (x)\sim(\underline{\quad\;}x\underline{\quad\;}\mathrm{sub}(\bar{g},\bar{g})\underline{\quad\;})$

whence by universal instantiation:

A6: $\vdash \sim(\underline{\quad\;}\bar{h}\underline{\quad\;}\mathrm{sub}(\bar{g},\bar{g})\underline{\quad\;})$

On the assumption that PA is consistent, A4 and A6 cannot both hold, and consequently $G$ cannot be a theorem.

## II. TRUTH AND PROVABILITY IN PEANO ARITHMETIC

If $G$ says that $G$ is not provable in PA, and if $G$ is in fact not a PA theorem, then it would appear that what $G$ says is true. $G$ would be true-but-unprovable. But this does not by itself show that human beings can know something—the truth of $G$—that Turing machines cannot know, since both the truth and the unprovability of $G$ are predicated upon the assumption that PA is consistent. If PA should turn out to be inconsistent, $G$ would be not true-and-unprovable, but provable-and-false.

There are therefore two alternatives: if PA is consistent, $G$ is unprovable and true; if PA is inconsistent, $G$ is provable and false. Whichever alternative holds, provability and truth part company. Either some truths are not theorems, or some theorems are not true. The domain of expertise of a Turing machine lies in the area of proof and provability, not in the area of truth. Human beings, on the other hand, are acquainted with both proof and truth, and also know of cases where the two diverge.

It might be argued, despite the above, that, if we conditionalize on PA's consistency, truth and provability come together again. Thus, on the assumption that PA is consistent, $G$ is true and unprovable, but at the same time the conditional sentence A1 is true and, as we shall see, provable in PA. Similarly:

A7: If PA is inconsistent, then $G$ is provable.

is also true and provable. If, it is claimed, we focus on conditional sentences, truth and provability coincide once more and Turing machines can know (=prove) whatever human beings can know (=prove).

The details of this claim must be examined. Denote the Gödel number (in Peano notation) of a well-formed formula $A$ by $[A]$, define the expression $\mathrm{Bew}[A]$, which says that $A$ is provable, as $(\exists x)(\_\_\_\_x\_\_\_\_[A]\_\_\_\_)$, and note that $G$ is equivalent to $\sim\mathrm{Bew}[G]$. The assertion that PA is consistent may be symbolized in PA by a number of equivalent expressions, such as $\sim\mathrm{Bew}[0 = 0']$. I abbreviate the latter by "Cons(PA)." Then A1 is written in the language of PA as:

A8: $\mathrm{Cons(PA)} \longrightarrow \sim\mathrm{Bew}[G]$

or equivalently as:

A9: $\mathrm{Cons(PA)} \longrightarrow G$

Proving A9 as a theorem of PA is tantamount to proving Gödel's second theorem, which states that Cons(PA) is not provable in PA unless PA is inconsistent. (If Cons(PA) *were* a PA theorem, then, by A9, $G$ would also be a theorem, which is known not to be the case.) In his original 1931 paper, Gödel does not prove this second theorem; a proof is to be found in David Hilbert and Paul Bernays.[4] Assuming that this proof can be fully formalized in PA, A9 is both true and provable in PA.

It follows from the above that in the case of the conditional sentences A8 and A9, truth and provability coincide, and a Turing machine can *know* the truth of A9 because it can prove A9 as a theorem. But a different result holds for the formal version of A2. This is:

A10: $\mathrm{Cons(PA)} \longrightarrow \sim\mathrm{Bew}[\sim G]$

---

[4] *Grundlagen der Mathematik, Volume 2* (Berlin: Springer, 1939). See Stephen Cole Kleene, *Introduction to Metamathematics* (Amsterdam: North-Holland, 1952), p. 211; and Elliott Mendelson, *Introduction to Mathematical Logic* (Princeton: Van Nostrand, 1964), p. 148.

Moving from A9 to A10 yields a candidate for the true-but-unprovable sentence that we seek.

Is A10 true? Suppose, for reductio, that PA is consistent and that $\sim G$ is a theorem. Then $[\sim G] \in Th$, and, from the consistency of PA we infer that $[G] \notin Th$. At the same time, since $\sim G$ is equivalent to $(\exists x)(\underline{\quad}x\underline{\quad}[G]\underline{\quad})$, from $\vdash \sim G$ we derive $\vdash (\exists x)(\underline{\quad}x\underline{\quad}[G]\underline{\quad})$. Because the open formula $(\exists x)\underline{\quad}x\underline{\quad}y\underline{\quad}$ weakly represents $Th$, it follows that $[G] \in Th$, which completes the reductio. Hence A10 is true.[5]

The informal argument just presented cannot be converted into a formal proof of A10's theoremhood, and there are good reasons to believe that A10 is in fact unprovable in PA. Because of the equivalence of $\sim G$ to $\mathrm{Bew}[G]$, A10 can be written as:

A11: $\mathrm{Cons(PA)} \longrightarrow \sim\mathrm{Bew}[\mathrm{Bew}[G]]$

In the logic of provability, one of the laws governing provability as a modal notion is:

A12: $\mathrm{Bew}[X] \longrightarrow \mathrm{Bew}[\mathrm{Bew}[X]]$

("If $X$ is provable, then it is provably provable").[6] Instances of A12 are theorems of PA. Since the converse of A12 does not hold, the consequent $\sim\mathrm{Bew}[\mathrm{Bew}[G]]$ of A11 is a net strengthening of the consequent $\sim\mathrm{Bew}[G]$ of A8, and it is unlikely that the long and complex proof of A8 in PA could be reworked so as to yield a proof of A11. Not impossible, perhaps, in the absence of a knock-down independence proof, but highly unlikely.

The upshot is that a Turing machine programed to run through proofs in PA will almost certainly never classify A10 as a theorem (unless, of course, PA is inconsistent), but neither will it be capable of assigning it to the special category of "unproven but true." Recognition of the existence of such a category therefore marks a difference in principle between human and machine thinking.

Despite the above reasoning, there will still be those who wish to maintain that somehow, in some fashion, a Turing machine can recognize the truth of A10. There are three possible strategies for arriving at this conclusion. The first is to show that "truth," like

---

[5] It is frequently stated that proving the nontheoremhood of $\sim G$ requires the assumption that PA is not only consistent but ω-consistent (see, for example, Mendelson, p. 143). The argument given in the text shows, however, that the weaker assumption of ordinary consistency suffices.

[6] See George Boolos, *The Logic of Provability* (New York: Cambridge, 1993), p. 15.

"provability," is in some sense expressible by a formula of PA. The second is to formalize the notion of truth by introducing a new primitive sentential operator *T*, the truth operator, and show that in PA, suitably enriched with extra axioms, *T*(A10) and similar theorems may be derived. The third is to hold that adding new axioms to PA, without the operator *T*, allows A10 to be proved, and that in stronger and stronger extensions of PA, a Turing machine can derive every proposition a human can know to be true. I shall argue that all these strategies fail.

## III. EXPRESSING TRUTH BY AN ARITHMETICAL FORMULA

Tarski's theorem states that the set *Tr* of Gödel numbers of well-formed formulas of PA which are true in the standard interpretation is not "arithmetical," that is, there is no open sentence ____$x$____ of PA such that, for any $x$, $x$ belongs to *Tr* if and only if the closed sentence ____$\bar{x}$____ is true in the standard interpretation.[7] The notion of "truth" in PA is consequently on quite a different footing from that of "provability." The latter concept is represented by an open arithmetical formula; no analogous formula expresses (much less represents) the former, thus reinforcing the hypothesis that "truth," though meaningful to humans is a closed book to a Turing machine.

## IV. INTRODUCING THE *T* OPERATOR

The second possible way of supplementing a Turing machine's ability to deal with proof by an ability to deal with truth is to introduce an explicit truth operator *T*. With suitable *T* axioms, the machine might then be able to prove propositions like:

A13: *T*(If PA is consistent, then $\sim G$ is not provable)

For steps in the implementation of this strategy, see, for example, Solomon Feferman.[8] There are, however, two worries. The first is that the *T* axioms may be able to do no better than supply, for any already-proved theorem *X*, a new theorem *T*(*X*). If *X* were itself not a theorem of PA, as is probably the case with A10, then *T*(*X*) would also fail to be a theorem. If this were so, introduction of the *T* operator would do nothing to correct a Turing machine's inability to cope with unprovable but true propositions.

The second worry is more general. *T* axioms may increase a Turing machine's deductive powers in the sense of giving it more theorems it can prove, that is, broadening the extension of "proof." But how is this relevant to "truth"? Take the analogy of non-Euclidean geometries. If a Turing machine is programed to accept the Gödel numbers of formal axiomatic proofs in Euclidean geometry, and if the parallel postulate is then replaced by Lobatchevsky's postulate and the machine is reprogramed, it will generate a different set of proofs and a different set of theorems. But will the reprogramed machine categorize the old parallel postulate as "true but unprovable"? Or as "formerly provable but false"? This seems unlikely. Between provability and truth there yawns an unbridgeable gulf.

---

[7] Alfred Tarski, "The Concept of Truth in Formalized Languages," in *Logic, Semantics, Metamathematics* (New York: Oxford 1956), pp. 152-278. See Mendelson, p. 151; Machover, p. 236.

[8] "Reflecting on Incompleteness," *Journal of Symbolic Logic*, LVI (1991): 1-49, here pp. 13 ff.

## V. STRONGER AND STRONGER EXTENSIONS OF PA

The third way of getting a Turing machine to accept A10 as true, without introducing the operator *T*, is to strengthen PA and deduce A10 from new axioms.

But again, there is a difficulty. No matter what system *L* of axioms and rules the machine is programed to use, and provided that *L* is at least as strong as PA so that "proof" and "substitution" are representable, there will always be a sentence analogous to A10 which is true, but for which the Turing machine will be unable to find a proof. In Zermelo-Frankel set theory, for example, an appropriately programed machine will be able to construct a Gödel sentence $G_{ZF}$ and a ZF-analogue of A10:

A14: $\mathrm{Cons}(ZF) \longrightarrow \sim\mathrm{Bew}[\sim G_{ZF}]$

The machine will be able neither to prove this sentence, nor to recognize it as true. A human being, on the other hand, can know that every such sentence, in every extension of PA, is true.

To sum up, what is the source of a Turing machine's inability to recognize the truth of A10? Can the inability be done away with by moving to more powerful programs? Let us start with PA.

A Turing machine, programed to run through and list all proofs valid according to the axioms and rules of PA, cannot recognize A10 to be true if it cannot prove it. For such a machine, there are only two categories of well-formed formulas: those which are theorems and those for which no proof has yet been found. (If PA is discovered to be inconsistent, the second category will turn out to be empty.) For the machine, the category "not yet proven but true" does not exist.

What now of stronger machines, equipped with more powerful axiomatic bases and able to generate larger sets of theorems? What of machines equipped with the truth operator *T*? What of machines with huge data banks? It is difficult to see why the situation, for

# COMMENTS AND CRITICISM

## DEFLATING THE CONSERVATIVENESS ARGUMENT

In "Proof and Truth: Through Thick and Thin," Stewart Shapiro[1] argues that deflationist theories of truth are not really so deflationist, in that they require a notion of logical consequence too strong to be codified in an effective deductive system. He gives two related arguments for this conclusion, and there are hints of a third; I think all of them are mistaken. I shall take no stand on whether we can understand a noneffective notion of consequence;[2] I claim only that the sorts of considerations Shapiro brings to bear do nothing to show a problem for a deflationist who sticks to first-order logic or some other effective logic.

The term 'deflationism' has been used to cover quite a few different approaches to truth, and it is not clear precisely what they should be taken to have in common. Fortunately, there will be no need for me to be very precise: I shall argue that on no reasonable construal of the term do Shapiro's arguments work.

Before presenting Shapiro's arguments, I shall call attention to one point to which every theorist of truth should agree: that by having a notion of truth we increase our expressive power in an important way. Indeed, this is a point that deflationists (or those who call themselves that) like to stress: the main point of having the notion of truth, many deflationists say, is that it allows us to make fertile generalizations we could not otherwise make; where by a *fertile* generalization I mean one that has an impact on claims *not involving the notion of truth*. If we disagree with everything we have ever heard the Pope say, we might conclude on inductive grounds that nothing the Pope says is true; this has the effect of committing us against not only what we have heard him say but also against what he has said which we have not heard. Since we do not know exactly what these unheard remarks are, there seems to be no way to disagree with all his remarks except by use of the notion of truth (or some other device to the same effect). That is the main point of the notion of truth, ac-

---

[1] This JOURNAL, XCV, 10 (October 1998). 493-521.

[2] My own view is that we probably can understand the noneffective notion of consequence involved in the full logic of the finiteness quantifier, but not the more powerful notion required for full second-order logic. Reasons for the differential treatment, and an explanation of why the conclusion can only be probable, can be found in my "Which Undecidable Sentences Have Determinate Truth Values?" in H. Garth Dales and Gianluigi Oliveri, eds., *Truth in Mathematics* (New York: Oxford, 1998), pp. 291-310.

0022-362X/99/9610/533–40

cording to many deflationists: it allows us to make commitments (about matters not involving truth) that we could not make without it. Shapiro says that deflationists hold that truth is "metaphysically thin." I am not sure what this means,[3] but one thing that it better *not* mean is that we cannot use it to express important things inexpressible or not easily expressible without it, or that we cannot use it to make commitments about matters not involving truth beyond those commitments which we could make without it; for it is a clear part of deflationist doctrine that truth is *not* metaphysically thin in *that* sense. (We might put this by saying that everyone, deflationist or not, must agree that truth is not *expressively thin*.[4])

Shapiro's arguments all turn on the well-known fact that in the framework of first-order logic, if you start out with a theory adequate to the elementary arithmetic of natural numbers (whether a theory of pure arithmetic or a more inclusive theory, say, one that postulates classes as well as numbers), and you add a truth predicate to the theory in a certain natural way, then the resulting theory $T^*$ is not a conservative extension over the theory $T$ from which one started.[5] To be slightly more precise, let us assume that we adopt a standard Gödel numbering of $T$, and that in $T$ we define denotation relative to that Gödel numbering in the usual way. We now assume:

(i) that for each of the finitely many atomic predicates $p$ of $T$, $T^*$ licenses the usual general rule about how the truth of any sentence of form $p(t_1, \ldots, t_n)$ depends on the denotations of $t_1, \ldots, t_n$; and that for each method of composing sentences out of simpler sentences, $T^*$ licenses the usual general rule about how the truth of the compound depends on the

---

[3] I do not mean to blame Shapiro for the unclarity of the phrase: he might reasonably say that the point of calling a theory of truth 'deflationist' is to say that truth is in *some* sense "metaphysically thin," and that it is up to the deflationist to specify the sense. He might also reasonably say that many people who call themselves deflationist have been less than clear about the kind of "thinness" they are attributing. I have no disagreement with this; but it is important not to attribute a sense of thinness that deflationists explicitly disavow. (In my own deflationist moods, I would say that the most fundamental notion of truth is a purely logical notion applicable only to sentences we understand, and serving solely as a device of generalization; all other notions of truth—for example, those applicable to other languages—are to be explained in terms of this, plus relatively modest auxiliary notions.)

[4] I concede that one deflationist, Paul Horwich—*Truth* (Cambridge: Blackwell, 1990)—has formulated his theory of truth in terms of principles that are not powerful enough to give 'true' the expressive role that we want it to have. He has been correctly criticized for this by Anil Gupta, "A Critique of Deflationism," *Philosophical Topics*, XXI (1993): 57-81. But Shapiro's criticism is not just of Horwich, and is rather different.

[5] The truth predicate need only apply to arithmetic sentences. If it does, and we take the language of $T$ to include numerals, we can simplify the truth theory by avoiding the notion of satisfaction. I tacitly suppose this done in what follows.

> truth of the components. (For these purposes, take a sentence whose main connective is a quantifier over the natural numbers as "composed out of" the infinitely many sentences resulting from dropping the quantifier and substituting a numeral for the quantified variable.)

Also we assume:

> (ii) that $T^*$ allows mathematical induction on formulas containing the truth predicate. ($T$ will, of course, allow inductions on all formulas in the old language not containing the truth predicate; but extending induction to formulas containing the truth predicate is an important addition.)

The well-known fact (a trivial consequence of Gödel's second incompleteness theorem) is that a theory $T^*$ that extends $T$ in this way cannot be a conservative extension of $T$, if $T$ is consistent: $T^*$ implies sentences in the language of $T$ that $T$ itself does not imply.[6] Shapiro claims that this fact makes trouble for a deflationist who wants to stick with first-order logic: supposedly, it forces the deflationist to adopt an expanded logic in which adding a truth predicate to our theories in the manner just outlined yields no new consequences in the old language. (An expanded logic with this feature must be very powerful: powerful enough to resist effective codification.)

As Shapiro concedes, it is easy enough to add a truth predicate to number theory *in a more limited way* and get a conservative extension of what we started with; indeed, there are several ways to do so. He mentions a way that does not even allow the generalizations in (i), but only allows the Tarski biconditionals (which are derivable from (i) but do not entail any generalizations); call the resulting theory $T_w$. But it is more interesting to add truth in a way that includes the general laws in (i), since I think it is clear that without such general laws the truth predicate would not serve its main purpose. Charles Parsons[7] has noted that, if in adding 'true' we just add the general laws in (i), the extension (call it $T_i$) is perfectly conservative (even in first-order logic); it is only when in addition one allows induction on predicates containing 'true' that we get new first-order consequences in the old language. Since truth can be added in ways that produce a conservative extension (even in first-order logic), there is no need to disagree with Shapiro when he says "conservativeness is essential to deflationism" (497). Shapiro's position, however, is that a deflationist must hold that adding 'true' to number theory *in the full-blooded way that involves (ii) as well as (i)* produces a conservative extension. Call this the *conservative extension assumption.* Since this assumption can only be so in a noneffective expansion of first-order logic,[8] the crucial question is: Is there any good reason for a deflationist to accept the conservative extension assumption? There are several distinguishable lines of argument for it in Shapiro's paper.

---

[6] The idea is simply that (i) and (ii) in conjunction with $T$ allow us to argue that all of the axioms of $T$ are true, and that the rules of inference of $T$ preserve truth, and therefore that all the theorems of $T$ are true. Also by (i) in conjunction with $T$, '0=1' is not true; so it is not a theorem of $T$. That is a sketch of how to derive in $T^*$ that '0=1' is not a theorem of $T$. But by the incompleteness theorem, we cannot derive this in $T$ itself (if $T$ is consistent), so $T^*$ does not conservatively extend $T$.

[7] For a proof, see his *Mathematics in Philosophy* (Ithaca: Cornell, 1983), pp. 213-15. The idea is that only predicative sets of natural numbers (and the obvious principles governing them) are needed to turn Tarski's inductive clauses for truth into an explicit definition from which the generalizations in $T_s$ follow; but such a predicative extension of $T$ is easily seen to be a conservative extension, so $T_s$ must be. (Even if you expand the induction schema in the original language into the claim that every set that includes 0 and is closed under successor contains all numbers, you still do not get the instances of the schema in the new language, absent impredicative assertions of set existence.)

Shapiro's simplest argument for the conservative extension assumption is as follows:

> If truth [yields new consequences in the old language when it] is admitted into the induction principle, it is counterintuitive to maintain that the notion is thin and insubstantial. How thin can the notion of arithmetic truth be, if by invoking it we can learn more about the natural numbers (499)?

But the way in which we "learn more about the natural numbers by invoking truth" is that in having that notion we can rigorously formulate a more powerful arithmetic theory than we could rigorously formulate before. There is nothing very special about truth here: using any other notion not expressible in the original language we can get new instances of induction, and in many cases these lead to nonconservative extensions. So the question becomes:

> How thin can the notion of arithmetic truth be, if by invoking it we can rigorously formulate a more powerful theory about the natural numbers?

But as I noted at the start of this paper, it is quite uncontroversial that the notion of truth can be used to make generalizations that cannot be made without it, and that these generalizations can be important in giving rise to commitments not involving the notion of truth. Not only is this generally recognized, it is often cited by deflationists as the main point of the notion of truth. Given this, it is hard to see any force in the question.[9]

---

[8] Here and elsewhere, I implicitly assume that contraction of first-order logic is not an option.

Shapiro's main argument is different, and cannot be dismissed merely by pointing out that no deflationist believes in "expressive thinness." That main argument (497-99) has two components. Let $B$ be a theory about a nonlogical subject matter, and let $B'$ be the result of adding 'true' to the language and extending $B$ by adding new axioms that are (in Shapiro's phrase) "essential to truth." Then if truth is deflationary, $B'$ should be a conservative extension of $B$: for $B'$ is the union of $B$ plus a truth theory TR; if it fails conservatively to extend $B$, then there is a sentence $\Phi$ in the truth-free language such that $B \cup \mathrm{TR}$ entails $\Phi$ but $B$ alone does not. In other words, the set of sentences $B \cup \{\neg\Phi\}$ is logically consistent and does not contain 'true', but TR is incompatible with it. Indeed, by compactness, TR must entail a 'true'-free sentence of form $B_1 \& \ldots \& B_n \supset \Phi$ that is not logically true (where the $B_i$s are in $B$). But surely (this part of the argument concludes), we should not call a notion of truth TR deflationary if a theory that involves only axioms that are "essential to it" has nontrivial (not logically true) consequences about matters not involving truth.

We can apply this to the case where $B$ is a theory $T$ adequate to number theory. Let us concede that not only are the Tarski biconditionals "essential to truth," so too are the general laws (i) used in so many applications of the notion of truth. If we concede only that, nothing of great interest follows: we already knew that the theory $T_t$ obtained by adding 'true' only with these laws is a conservative extension of $T$. The potentially interesting result comes from assuming that the induction axioms involving true are also "essential to truth": since the result $T^*$ of adding these to $T_t$ is not a conservative extension of $T$ if the logic is effectively codifiable, we get:

(1) If the new induction axioms $J$ involving 'true' are essential to truth, and logic is effectively codifiable, then the notion of truth is substantial (not deflationary).

---

[9] I might also remark that since deflationists invariably agree that there can be fertile generalizations that involve the notion of truth, there is no obvious reason why a deflationist cannot allow such generalizations to be used in explanations. I think that any view reasonably called 'deflationist' must pose limits on the *ways* in which the notion of truth (and of truth conditions) is put to use in explanations, but I would have thought that any use of 'true' in explanations which derives solely from its role as a device of generalization should be perfectly acceptable. Shapiro makes a big deal of the fact that the notion of truth can be used in explanations by virtue of its generalizing role (505-07), but this fact seems to me quite unsurprising and no threat to any reasonable version of deflationism.

That is the first component of Shapiro's main argument.

The second component of Shapiro's main argument emphasizes the importance of the induction axioms: they are important, for instance, because we certainly want to be able to prove inductively that all theorems of $T$ are true on the basis of the truth of the axioms and the truth preservation of the rules of inference, and the axioms $J$ are obviously needed for such a derivation.[10] So

(2) The new induction axioms are essential to truth.

From (1) and (2) it follows that, if logic is effectively codifiable, then truth is substantial.

But I think it is clear that this argument rests on an equivocation with the phrase 'essential to truth'. What the argument for (2) establishes is:

(2′) The induction axioms are needed if we are to arithmetically derive important facts that involve the notion of truth.

But the argument for (1) works only if its first antecedent is interpreted as meaning:

(3) The truth of the induction axioms depends only on the nature of truth.

(If the truth of these axioms depend on facts about numbers as well, it is hardly surprising that number-theoretic conclusions should follow from them in conjunction with the purely truth-theoretic axioms $T_1$!) But (3) clearly does not follow from (2′). Moreover, (3) just seems false: the corresponding axioms would hold for any other predicate, and what they depend on is a fact about the natural numbers, namely, that they are linearly ordered with each element having only finitely many predecessors.

I have discussed the two most explicit arguments in Shapiro's paper. But there are two points which Shapiro emphasizes on which I have put no weight; perhaps they are the basis for a third argument. The first point is that mathematical theories adequate to number theory usually are formulated in terms of axiom schemas (that is, it is laid down that all sentences of a certain form are axioms), and on a typical such formulation, the additional induction axioms *J* result simply from extending the range of instances of the schema to the larger language that includes 'true'.[11] The second point concerns what is normally involved in commitment to a schema, at least in mathematics: the point is that people who commit themselves to the arithmetic induction schema mean to be committing themselves not only to its instances in the language they have but to all instances in any legitimate expansion, including one with a truth predicate; and similarly for other mathematical schemas, such as schemas of set existence.[12]

---

[10] It is worth noting that though we might also hope to establish that all theorems *of* $T^*$ are true, there is no hope of doing this in $T^*$: for $T^*$ does not have the means for attributing or denying truth to the axioms about truth in $T^*$. More generally, Gödel's second incompleteness theorem implies that (in a language adequate to number theory) no consistent effective theory of truth that allows us to pass freely between the attribution of truth to a sentence and the sentence itself can have the means for inductively arguing that all its theorems are true.

I fully concede both points, but they lend no credence to the conservative extension assumption, nor do they in any other way

---

[11] This is not so for all mathematical theories adequate to arithmetic. For instance, consider a theory of natural numbers and predicative classes of them. This can be formulated nonschematically, with only finitely axioms, by formulating induction as a single axiom using classes and axiomatizing the classes nonschematically in the manner of Gödel-Bernays set theory. Adding 'true' to such a theory requires new class-existence assumptions or new instances of induction that obviously are *not* obtained by extending an old axiom schema. And that certainly suggests that the extension of schemas is not central to the topic of the conservativeness of the truth predicate: a conclusion soon to be argued on independent grounds.

[12] This second point need not be so for schematic theories outside mathematics. Shapiro notes the possibility of simulating arithmetic within physical theory: perhaps by singling out a specific physical object to correspond to 0, and defining a predicate *N* (corresponding to 'number') true only of certain physical objects located on a ray emanating from this 0, and certain operations on the *N*-objects corresponding to successor, addition, and multiplication. To adopt the analogues of the first-order arithmetic axioms is to adopt a certain theory of the *N*-objects: one which is committed to the set of *N*-objects being infinite in extent, and which is also committed to induction holding on *N*-objects, for predicates in the language of the theory. Suppose that *N* is so defined that the successor of an *N*-object is always at least one mile from that object; then, if we assume that space is Archimedean (no two objects are infinitely far from each other), induction must also hold for predicates not expressible in the language. But we can imagine the possibility of adopting a cosmology on which space is non-Archimedean; or at least, of leaving the possibility of such a cosmology open. In that case, it would not be out of the question to accept induction on *N*-objects for predicates in the original language, but reject or remain noncommital about accepting it for predicates in the expanded language. (For *N* defined in a way that imposes no distance requirement on the successor of an object, we might be able to get an example that did not require our theorist to take non-Archimedean space seriously.)

The analogous attitude for natural numbers is, of course, absurd: it is part of the idea of natural numbers that they are Archimedean, that is, it is part of the idea that induction holds on any predicate of numbers we could ever come up with. *It is something about our idea of natural numbers* that makes it absurd to suppose that induction on the natural numbers might fail in a language expanded to include new predicates (whether truth predicates or predicates of any other kind): nothing about truth is involved make trouble for a deflationist who wants to stick to first-order logic. In the first place, it is hard to see how the fact that (in mathematics at least) we are committed in advance to extend our schemas to new predicates could possibly show that the theories so extended conservatively extend the theories prior to the extension. In the second place, this point about the extension of schemas has nothing specially to do with truth: we are committed in advance to extending our schemas to all new predicates, not just 'true'.

Late in the article (511) Shapiro seems to suggest that the fact that we are committed in advance to extending induction to any new predicates we introduce shows that our understanding of the natural numbers exceeds the bounds of first-order logic. I think this is a mistake, but shall not discuss it here. There is no need to, because if it is correct it gives reason to go beyond first-order logic which is quite independent of deflationary theories of truth. The commitment to extend induction to newly introduced predicates has nothing special to do with truth: it applies to all predicates, not just truth. So it is hard to see how one's theory of truth could have any special relevance to it.[13]

HARTRY FIELD

New York University

---

[13] Since it is only the expansion of the schema to new instances involving 'true' that Shapiro makes use of, we might be able to avoid all issues of indefinite extensibility by imagining that we start out by accepting a schema of induction *in a language that contains its own truth predicate* This would not actually avoid the issue of indefinite extensibility if, like Tarski, one takes the position that there is an infinite hierarchy of truth predicates for a given language. for then adding any one yields a nonconservative extension of the original language. And Shapiro does seem to hold to the Tarskian view. But there is a respectable alternative position: that we have a single truth predicate for our language, applicable to sentences that contain it as well as sentences that do not. Of course, we know from Tarski that the truth theory in such a language cannot be fully classical, in the sense that not all instances of the Tarski schema '"*p*" is true if and only if *p*' can be accepted. (More accurately, a nonclassical alternative must either restrict the Tarski schema *or restrict classical logic*; I shall not consider the latter possibility here.) For an exploration of an important family of nonclassical alternatives (that keep classical logic), see Harvey Friedman and Michael Sheard, "An Axiomatic Approach to Self-referential Truth," *Annals of Pure and Applied Logic*, XXXIII (1987): 1-21. And for a discussion of how this connects up to the point of footnote 10, see the last section of my "Which Undecidable Sentences Have Determinate Truth Values?"

# COMMENTS ON SHAPIRO*

Let us say I am committed to deflationist truth: I think augmenting a language with a truth predicate, Alfred Tarski's perhaps, or some substitutional variant, does not saddle me with new metaphysical commitments: "truth is not a metaphysically substantial property, and it has no nature,"[1] I am prone to say (whenever the topic of truth comes up).

Stewart Shapiro challenges the idea that this sits comfortably with "the logician's use of truth as a sign of expressive and proof-theoretic *strength*," and hopes to show the only way for the deflationist to resolve this tension is to embrace a "very strong, noneffective notion of logical consequence" (494-95). Since the aim is to show the deflationist is forced to a noneffective consequence relation, let us keep our eyes firmly on Shapiro's target, the deflationist commited to the first-order consequence relation, what I shall call the *first-order deflationist*.

Shapiro's ingenious argument is quite intricate, but we need only concern ourselves with the considerations that he takes to give rise to the tension with which the first-order deflationist must deal. As with Shapiro, let $\mathcal{L}$ be a language whose syntax is effective, and suppose $\mathcal{L}$ can express a certain amount of arithmetic. The variables of $\mathcal{L}$ range over the natural numbers, and we introduce a Gödel numbering: $\ulcorner\Phi\urcorner$ which assigns every natural number a sentence.[2] $A$ is an effective and sound theory in $\mathcal{L}$ and contains the rudimentary axioms of arithmetic. $\mathrm{PRF}(x, y)$ is the proof predicate for $A$ definable in $\mathcal{L}$, and let $G \equiv \neg\exists y\mathrm{PRF}(\ulcorner G\urcorner, y)$, the fixed point for $\neg\exists y\mathrm{PRF}(x, y)$.

Shapiro concedes that a new system $A'$ containing a truth predicate (following Tarski) can be constructed, one which is capable of proving every T-sentence and is conservative with respect to $A$. Shapiro claims, however, that such a theory does not contain everything "essential" to "arithmetic truth" (499). What is required of "an adequate theory of truth" is that it have "the resources to *establish*" that the axioms of $A$ are true and that the rules of inference preserve truth. In short, it should be able to establish:

(*) $\forall x(\exists y\mathrm{PRF}(x, y)\rightarrow T(x)$

---

* My thanks to Jonathan Kastin, Arnold Koslow, and George Smith for comments

[1] Stewart Shapiro, "Proof and Truth: Through Thick and Thin," this JOURNAL, XCV, 10 (October 1998): 493-521, here p. 494.

[2] Shapiro chooses a Gödel numbering onto the natural numbers, and I follow suit.

Furthermore, it should also be able to establish $\neg 0=1$, and thus, via a T-sentence, $\neg T(\ulcorner 0=1 \urcorner)$. But under such circumstances, it can establish:

$$\neg\exists y \mathrm{PRF}(\ulcorner 0=1 \urcorner, y)$$

which says that $A$ is consistent, and therefore is not provable in $A$ (by Gödel's second incompleteness theorem). So this additional "essential" element of the theory of truth yields a theory of truth for $A$ which is not conservative (with respect to $A$).

Something slippery just happened here. Although the deflationist agrees that "generalizations" concerning truth *are* "essential to" a theory of truth (and by this, she means, for example, clauses such as, $\ulcorner \Phi \urcorner$ is true and $\ulcorner \Psi \urcorner$ is true if and only if $\ulcorner \Phi \,\&\, \Psi \urcorner$ is true), she does not think it "essential to" a theory of truth to establish generalizations about *truths*. What *is* true (and how) is not, properly speaking, part of the *theory* of truth. In particular, although the deflationist is perfectly willing to describe $\neg 0=1$ and (*) as perfectly obvious, she still has no reason whatsoever to regard *them* as part of the theory of truth. Call *physical truth* the truths in some physical language $\mathcal{L}_P$ that follow from some physical theory $A_P$: surely, it is not required of the deflationist that when she supplements $\mathcal{L}_P$ with a theory of truth that it should follow from her theory of truth that everything that follows from $A_P$ is true. And if she does introduce the theory in such a way that this can be established from her theory of truth, why should she expect the resulting theory to be conservative (and, anyway, why would she call such a theory thus augmented a "deflationist" one)?

The point is a simple one: the capacity to establish (nonlogical!) truths and generalizations about such goes quite beyond what a first-order deflationist calls a deflationist theory of truth—and this regardless of how obvious such truths and generalizations (about them) are. The first-order deflationist can even grant that in the case of a theory of truth for arithmetic it is desirable to introduce resources for *characterizing truths* that go beyond deflationist truth ((*), for example): in doing so, of course, she notes that the resulting bundle is not *deflationist* truth and so one should hardly be surprised at its nonconservativeness.

The second argument Shapiro offers that the first-order deflationist is involved in a tension involves the induction scheme for arithmetic. Given "something like" Tarskian satisfaction in $A'$, if we allow instances containing the truth predicate $T$ in the induction scheme, then the resulting truth theory is not conservative (499).

The first-order deflationist, in light of what has been said above, should simply deny that allowing instances containing $T$ in the induction schema results in a *deflationary* truth theory and, of course, then there is no surprise at the resulting theory failing to be conservative. But Shapiro invokes Michael Dummett[3] (and Shaughan Lavine) as authorities against this move, because "it is part of the concept of natural number, *as we now understand it*, that induction with respect to any well-defined property is a ground for asserting all natural numbers to have that property" (500; italics added by Shapiro). What follows is an argument that, if someone admits that a predicate $P$ (however defined) "has a determinate extension over the natural numbers (or whatever the numerals denote)," then he must allow that predicate to appear in instances of the induction scheme (501).

Unfortunately, for this line of thought, these intuitive considerations apply *only* to the *standard model* of arithmetic, and only to concepts of natural number so restricted: it is simply *not true* of other models of $A$ that any predicate $P$ (however defined) with a determinate extension over the numbers of that model belongs in the induction scheme. Indeed, what allows the existence of nonstandard models for any first-order theory containing arithmetic in the first place is the fact that although the induction schema holds of any predicate definable in the language of that theory (and so the induction schema restricted to predicates of that language holds in *all* models of that theory) this is not true of every definable predicate in that same language enriched with a truth predicate; not all of the models of the restricted induction schema are models of the induction schema relative to the new language. This is why allowing the truth predicate in the induction schema is not conservative. The *first-order* deflationist (especially that deflationist who has conceded to Shapiro that "metaphysical robustness lines up with logical consequence" (498)[4]) will therefore deny that $A$ picks out the standard model to the exclusion of the other models in which $A$ holds (and therefore will deny Shapiro's claim that, "The idea is that the notion of arithmetic truth is somehow implicit in the arithmetic concepts [and] [t]his much is congenial with the main, but vague deflationist theme that the truth predicate *T* refers to an insubstantial property" (501). Arithmetic truth here refers, as before, to truth in the standard model, and the first-order deflationist will deny that *A* (or *any* effective axiomatization of arithmetic for that matter) somehow (!) implicitly grasps truth in the standard model except in the mere sense that the standard model is implicitly grasped among other nonstandard models which are also implicitly grasped. Shapiro's argument can trap the first-order deflationist into conceding a noneffective consequence relation only if she already *has* implicitly adopted the idea that somehow a first-order axiomatization itself noneffectively (but implicitly) picks out the standard model.

---

[3] "Reply to Wright," in Brian McGuinness and Gianluigi Oliveri, eds., *The Philosophy of Michael Dummett* (Dordrecht: Kluwer, 1994), pp. 329-38, cited in Shapiro.

[4] The point of 'especially' is that a deflationist need not even concede that metaphysical robustness line up with logical consequence. She may be willing, for example, to tie together an effectively characterized sequence of theories in language containing truth predicates for the theories earlier in the sequence, and describe all such theories (stipulatively) as about the same subject matter even though they are *not* conservative with respect to each other. Of course, such a set of (first-order) theories no more "grasps" the standard model than any first-order theory does.

I submit, therefore, that Shapiro has failed to show any tension that a first-order deflationist need feel between a commitment to deflationist truth as described above, and the fact that—in certain contexts—a Tarskian truth predicate is not conservative.

JODY AZZOUNI

Tufts University

THE JOURNAL OF PHILOSOPHY

VOLUME XCVI, NO. 11, NOVEMBER 1999

# MEANING AND DEFLATIONARY TRUTH*

"Charity...is not puffed up"—I Corinthians, 13, 4.

I would like to discuss Donald Davidson's[1] reasons for rejecting deflationary views of truth, while being particularly concerned with his claim that deflationism must be rejected because the concept of truth plays an essential explanatory role in the theory of meaning. Accordingly, my discussion will bear on the more general question as to what deflationists about truth can or should say about meaning.

## I INTRODUCTION

Davidson agrees with the deflationists in opposing all traditional theories of truth.[2] He denies, however, that this rejection requires replacing accounts of truth in terms of "correspondence" or "coherence" with something more stripped down. The deep mistake, which deflationists share with traditionalists, is to define truth at all. We can shed light on the concept of truth by tracing its relations with other important concepts: notably, belief and meaning. But we cannot define it. At the end of the day, we have to accept the concept of truth as an *exploratory* primitive.

This hostility to deflationary accounts of truth ought to strike us as puzzling. Davidson's views about meaning owe large and gener-

---

* I want to thank Paul Horwich and Hartry Field for numerous extremely helpful discussions of the issues dealt with in this paper

[1] "The Structure and Content of Truth," this JOURNAL, LXXXVII, 6 (June 1990): 281-328, "The Folly of Trying to Define Truth," this JOURNAL, XCIII, 6 (June 1996): 263-78.

[2] For this reason, Richard Rorty has tried to recruit Davidson for the deflationist camp. See Rorty's "Pragmatism, Davidson, and Truth," in Ernest Lepore, ed., *Truth and Interpretation* (New York: Blackwell, 1986), pp. 333-68 He does not distinguish clearly, however, between two ways in which there might not be "much to say about truth". adopting a deflationary account and taking truth as a primitive.

0022-362X/99/9611/545–64

UNIVERSITY LIBRARY
ALIPUR
CAMPUS

ously acknowledged debts to W. V. Quine, who is a deflationist with respect to truth. To be sure, Davidson's views diverge from Quine's in important respects. But it is far from obvious that the differences mandate a wholly distinct attitude to truth. In my view, they do not.

## II TRUTH: FUNCTION AND MECHANISM

Let us glance at some current deflationary approaches to truth, starting with Quine's[3] disquotational theory. According to Quine, the meaning of 'true' is fixed by certain logical equivalences. Thus:

> 'All that glisters is not gold' is true if and only if all that glisters is not gold; 'France is hexagonal' is true if and only if France is hexagonal... and so on.

Appending 'is true' to a quoted sentence is just like canceling the quotation marks ("disquotation"). For this approach, the core use of 'true' involves cases where a sentence is referred to by its quotation-mark name. But given additional empirical information—for example, that a certain sentence expresses Fermat's Last Theorem—we can easily ascribe truth to sentences picked out in other ways.

The disquotational theory treats 'true' as a predicate of sentences. But similar accounts of the behavior of the truth predicate can be given, mutatis mutandis, if we prefer to think of 'true' as predicated of propositions. The axioms of such a theory will be:

> The proposition that all that glisters is not gold is true if and only if all that glisters is not gold...and so on.

Paul Horwich's[4] "minimal theory" takes this form.

A recent variant of disquotationalism is Robert Brandom's[5] version of the "prosentential" account of truth. According to Brandom, we should see '...is true' not as a predicative expression but as a sui generis prosentence-forming operator. Prosentences stand in for sentences in the way that pronouns stand in for referring expressions. Thus, appending '...is true' to a sentence nominalization yields a prosentence which, as a whole, inherits its content anaphorically from whatever sentence tokening its contained sentence nominalization picks out (*ibid.*, pp. 299-305).

---

[3] *Pursuit of Truth* (Cambridge: Harvard, 1990).

[4] *Truth* (Cambridge: Blackwell, 1990).

[5] *Making It Explicit* (Cambridge: Harvard, 1994).

Compare any of these accounts of truth with Frank Ramsey's[6] redundancy theory, according to which 'It is true that Caesar was murdered' means no more than that Caesar was murdered (*ibid.*, pp. 34-35). (Any difference is entirely "stylistic": for example, we may use 'It is true that...' to speak more emphatically.) In effect, the redundancy theory treats 'It is true that...' as a sentential connective with the same truth table as double negation. As a result, the redundancy theory cannot handle cases where is '...is true' seems to function as a predicate attached to sentences designated by a singular term, as in 'Fermat's Last Theorem is true'. To cope with these cases, truth talk must be seen as involving some kind of denominalizing operation. This is just what contemporary versions of deflationism let us do.

The great advance of Quine's view over the redundancy theory, then—an advance which its current rivals preserve—is that it shows how truth talk contributes significantly to the expressive resources of our language. For a disquotationalist, 'true' offers a way of replacing talk about the world with logically equivalent talk about words. This move to the level of talk about words ("semantic ascent") gives us new things to generalize over—that is, linguistic objects, sentences—thereby enabling us to express agreement and disagreement with sentences that we cannot specify: for example, because we do not know exactly what they are ('What the President said is true') or because there are too many of them ('Every sentence of the form "*P* or not *P*" is true').

We see from this that we must distinguish two ways in which deflationary approaches to truth aim to be "deflationary." The first—and fundamental—has to do with the function or significance of truth talk. On this issue, contemporary deflationists are in broad agreement: the function of truth talk is *wholly* expressive, thus *never* explanatory. As a device for semantic assent, the truth predicate allows us to endorse or reject sentences (or propositions) that we cannot simply assert, adding significantly to the expressive resources of our language. Of course, proponents of traditional theories of truth do not deny any of this. What makes deflationary views deflationary is their insistence that the importance of truth talk is *exhausted* by its expressive function.

The second aspect of deflationism concerns the mechanism of truth talk: the logico-linguistic means by which truth talk accomplishes its expressive function. Here deflationists diverge: Quine speaks of disquotation, Horwich treats truth as a predicate of propositions, Brandom thinks of it as a sui generis prosentence-forming operator, and so on.[7] Nevertheless, there remains a common intention to explain the workings of truth talk in a metaphysically parsimonious way. So from a theoretical standpoint, these competing suggestions are all deflationary.

---

[6] "Facts and Propositions," *Proceedings of the Aristotelian Society*, Supplementary Volume VII (1927): 153-70, reprinted in D. H. Mellor, ed., *F P Ramsey: Philosophical Papers* (New York: Cambridge, 1990), pp. 34-51.

These two aspects of deflationism—functional and theoretical—are obviously connected. Traditional theories of truth—correspondence, coherence, pragmatist—presuppose that truth is an explanatorily significant property. Just as, because acidity is an important explanatory property in our accounts of chemical reactions, we want also to explain what it is that makes some substances acidic, so a philosopher who supposes that truth has important explanatory work to do—in the philosophy of language or epistemology—will want to explain what makes true sentences (propositions, beliefs, and the like) true. By contrast, if we regard the function of truth talk as only expressive, there need not be any complicated story about the nature of truth. Given deflationism with respect to truth talk's function, it makes sense to suppose that its mechanism can be explained using fairly limited theoretical resources.

## III. IS THERE MORE TO TRUTH THAN CONVENTION-T?

Davidson takes it to be common ground among deflationists that Alfred Tarski[8] has told us pretty much everything we need to know about truth. In Davidson's view, however, Tarski does not really define *truth* because his method for defining "truth predicates" for particular formal languages does not explain what those predicates have in common. Several ideas are in play here.

As is well known, Tarski declines to define a general notion of truth because he thinks that such a notion invites paradox. Instead, he shows how to define, in a suitable metalanguage, a metalinguistic predicate 'true-in-*L*' for an appropriate language *L* (the object language). Since he defines truth in terms of satisfaction, a generalized notion of reference, from a deflationary standpoint, Tarski might be regarded as telling us more than we need to know about truth. Still, his constructions are deflationary in spirit because he defines satisfaction (for *L*) by cases. And this is Davidson's point: such a procedure offers no general understanding of anything since "a definition given in this way can provide no clue for the next or general case." This limitation is inherited by the truth definitions themselves in that "Nothing in Tarski's truth definitions hints at what it is that these definitions have in common." We must conclude, therefore, that "there is more to the concept of truth—something absolutely basic, in fact—which Tarski's definitions do not touch."[9]

---

[7] There may be less to these differences than meets the eye.

[8] "The Concept of Truth in Formalised Languages," in *Logic, Semantics, Metamathematics* (New York: Oxford, 1956), pp. 152-278.

None of this is an objection to *Tarski* and Davidson does not suppose that it is. Tarski's truth predicates are defined for certain formal languages, where the interpretation of basic terms is a matter of stipulation. If we want to extend such a language by adding a new term, we can assign it whatever extension we choose. We do not need any "clues" as to how to go on. But to say this is simply to point out the gap between Tarski's constructions and ordinary truth talk.

Now, it is true that, for Tarski, when we extend a language by adding new terms, we introduce a wholly new satisfaction relation, and so a wholly new truth predicate. But it is misleading to say that Tarski does not even hint at what this new predicate has in common with the old. What makes all Tarski's predicates *truth* predicates is that they meet the same standard of material adequacy. The standard of adequacy in question is, of course, the famous *convention-T*:

(T) *X* is true if and only if *p*.

(where '*X*' is a name for some sentence of the object language and '*p*' is a sentence of the metalanguage that correctly translates *X*). A definition of truth will be considered "materially adequate" only if it allows us to prove each member of the appropriate set of equivalences (if it "meets convention-T"). Of course, the formal truth definitions do not themselves make this clear: that is, they do not themselves say why the predicates they define are truth predicates. Nor do they explain the point of having such metalinguistic predicates. But why is this a problem?

As Tarski makes clear, and as Davidson is of course well aware, it is via convention-T that his formally defined predicates make connection with our intuitive conception of truth. I think that this justifies Horwich's claim that equivalences of the sort to which convention-T points, and not Tarski's constructions, are what really give the content of the intuitive notion of truth. We may therefore want to say—particularly in the light of their restriction to artificial formalized languages and their enumerative approach to defining satisfaction relations—that these constructions add nothing to whatever understanding of truth is implicit in convention-T. It is only in virtue of that prior understanding that we recognize Tarski-style definitions as having anything to do with truth. But that *is* a comment on Tarski, not on deflationism generally. It does not show that the deflationist account of what constitutes that prior understanding is inadequate.

---

[9] "The Structure and Content of Truth," pp 287, 288

The suspicion may still linger that convention-T does not really explain what such definitions have in common. This suspicion is justified. A criterion of adequacy will not itself explain the point of having metalinguistic predicates that satisfy it. But as we have seen, deflationism involves an account of the function of truth talk as well as (rival) accounts of its mechanism. This is the point I once tried to make by responding to Davidson's question "What do such predicates have in common?" with the remark: "the utility, for any language, of having its own disquotation device." Davidson comments that it is difficult to assign a clear meaning to the "utility" of a device. But the utility in question is simply the addition to our expressive resources which comes with a device for semantic ascent.[10]

So far, then, Davidson's reflections on convention-T fail to pose significant problems for deflationary approaches to truth. He has a much more serious objection to make, however. A further reason for rejecting the idea that convention-T goes to the heart of our concept of truth is that:

> in those cases where the object language is contained in the metalanguage, the requirement is merely syntactical: it tells us something about the predicates, but not much about the concept. In other cases its application depends on our prior understanding of the notion of translation, a concept far more obscure than that of truth.[11]

Consider Quine. He explains the core use of 'true' by appeal to equivalences that involve sentences and their quotation-mark names. Following Hartry Field,[12] we can call the concept of truth thus defined *purely disquotational*. Pure disquotationalism does characterize the mechanism of truth syntactically. Evidently, this austere notion of truth, as grasped by a given person, applies only to utterances that the person in question understands. Equally evidently, however, ordinary truth talk ranges more widely than this: hence the standard objection to disquotationalism, of which Davidson's point is a version, that we sometimes wonder about the truth of an utterance we do not understand. Theoretical austerity, it seems, compromises functional adequacy.

---

[10] "The Structure and Content of Truth," p. 296, note 33.

[11] *Ibid.*, p 296

[12] "Disquotationalism and Factually Defective Discourse," *Philosophical Review*, CIII, 3 (July 1994): 405-52. See especially pp. 405-09.

This objection is not immediately decisive. As Field points out, a deflationist can say that, in the sort of case Davidson finds problematic, we are "wondering whether a good translation of the utterance will map it into a disquotationally true sentence we do understand" (*ibid.*, p. 408). Nevertheless, the objection poses a significant explanatory problem for the disquotationalist. If functional adequacy is not to compromise theoretical parsimony, whatever extensions of truth talk such examples require must be explicable in terms of disquotational truth together with fairly limited additional resources.

On this issue, I see no serious difference between Horwich and Quine. They may appear to diverge, since Horwich's minimal theory invokes propositions. If he can unpack talk of propositions in a suitably parsimonious way, however, the difference will be more apparent than real. Once we admit the need for extended disquotationalism, we see that any deflationist will have to say something about meaning; and if he wishes to avoid compromising his theoretical deflationism, he will have to be careful what he says.

Careful in what way? Again following Field, we can agree that the more radical and interesting forms of deflationism will not count objective interpersonal synonymy as belonging to these resources. Thus, a deflationist who appeals to a notion like "good translation" will add that "'good translation' should be taken to be a highly context-sensitive and interest-relative notion; it should not be understood as meaning 'correct translation', for talk of correctness of interpersonal translation makes no sense" (*ibid.*).

But Field may be understating the antideflationist challenge. With respect to extending disquotational truth in a way that remains theoretically deflationary, the problem is not just one of eschewing appeal to objective synonymies. The real disaster would be if our notion of good translation had to be explained in terms of a nondisquotational notion of truth. This would be a disaster no matter how context sensitive our notion of translation turned out to be. If meaning (or translation) depends on truth, and if an adequate functional deflationism requires extended disquotationalism, the attempt to be a functional deflationist will be self-defeating. This is just what Davidson thinks. Because he explains meaning in terms of truth conditions, he cannot be a deflationist—functionally or theoretically—with respect to truth.

In saying that he accepts truth as a primitive notion, Davidson means that the concept of truth is not something we should try to define or, if we doubt the possibility of giving a strict definition, capture by some "brief criterion, schema [or] partial but leading hint."[13] He does not mean that the concept cannot be illuminated. But illumination is to be found in tracing "the connections between the concept of truth and the human attitudes and acts that give it body" (*ibid.*). Davidson has in mind believing, asserting, interpreting, understanding, and so on.

Deflationism, as Davidson understands it, takes the opposite view. It reflects just the urge he thinks we must resist: to characterize truth exhaustively in terms of some brief formula. Accordingly, deflationism is committed to the view that the concept of truth can be understood in isolation from such concepts as belief or meaning. But although "[t]he disquotational feature of truth, wedded to the thought that this may exhaust the content of the concept of truth, encourages the idea that truth and meaning can be kept quite separate" (*ibid.*, p. 271), this idea is mistaken. With such concepts as belief, truth, and meaning, none is to be defined. But all can be illuminated by tracing their relations with each other.

Now, while Horwich does remark that truth has "a certain purity,"[14] I am not sure that it is correct to represent deflationists as holding that truth can be understood *fully* in *complete* isolation from considerations having to do with meaning, belief, and the like. As we have seen, disquotationalists accept the need for an extended notion of disquotational truth. All they are committed to is explaining the extension in suitably deflationary terms: in particular, without covert appeal to some richer-than-disquotational truth concept. So the question is: Does Davidson's approach to meaning give us reason to think that they must fail in this attempt? I do not think that it does.

## IV THEORIES OF MEANING

The key to understanding why Davidson's hostility to deflationism is misplaced is to distinguish two senses of the phrase 'theory of meaning'. In a broad sense, a philosopher's theory of meaning is his account of what meaning consists in. But although Davidson certainly advances a theory of meaning in this sense, he also uses the phrase 'theory of meaning' more narrowly. In this narrow sense, a Davidsonian theory of meaning is an axiomatic theory that generates the meaning of every sentence of some target language. Now, since the 'meaning' thus generated is simply the sentence's truth condition, it is natural to suppose that Davidson identifies meanings with truth conditions, so that his broad sense theory of meaning is that meanings are truth conditions. If this were so, his conception of a (narrow sense) theory of meaning would determine his general account of what meaning is, so there would not be much point in insisting on the distinction. But, of course, this cannot be the end of the story. Davidson owes (and provides) an account of why a device for generating truth conditions should be thought to specify meanings, as well as an account of the adequacy criteria to which such a theory of meaning is subject. His theory of meaning (broad sense) involves more than a simple identification of meanings with truth conditions, if indeed it involves such an identification at all.

---

[13] "The Folly of Trying to Define Truth," p. 276.

[14] Horwich, quoted by Davidson, "The Folly of Trying to Define Truth," p. 274.

Davidson's split-level approach to meaning reflects his acceptance of Quine's skepticism about reified "meanings." For Quine, since there are no meanings—in the head or anywhere else—meaning is just whatever translation preserves. We therefore achieve an exhaustive understanding of meaning if we determine (a) what needs to go into an adequate translation manual between our own language and some alien tongue and (b) the constraints that guide and govern the construction of such manuals. Quine thinks that the constraints on translation fall far short of enforcing a unique choice of manual. Translation—hence meaning—is indeterminate.

Davidson prefers "interpretation" to "translation"; and his methodological views diverge in some ways from those of Quine. But *strategically* his approach to meaning is the same. If we (a) determine the theoretical form of the knowledge at which interpretation aims and (b) fix the methodology of interpretation, then we have said all there is to say about meaning. For Davidson, meaning just *is* whatever the practice of interpretation reveals.

We can now see more clearly why, when thinking about Davidson, we must be careful not to conflate the two uses of the term 'theory of meaning'. Using the term narrowly, as Davidson himself often does, we refer to some axiomatic theory, a recursive device for specifying the meaning of every sentence of a given language or, more precisely, of the current idiolect of a particular speaker. No such particular theory constitutes an account of what meaning *consists in*, however, any more than Tarski's constructions tell us why the metalinguistic predicates they define are truth predicates. To determine Davidson's views about the nature of meaning, we must attend to his ideas about the constraints that particular theories of meaning must satisfy to be judged acceptable: his theory of interpretation. For clarity, I shall henceforth use the phrase *theory of meaning* exclusively in the narrow sense. I shall refer to Davidson's more general, interpretationist conception of meaning as his *theory of interpretation.*

Making this distinction dispels the illusion that Davidson "explains meaning in terms of truth conditions." Certainly, he uses truth talk in (narrow sense) theories of meaning. But the use is expressive: truth talk is used in the course of *specifying* particular meanings. Davidson would *explain* meaning in terms of truth only if he needed to appeal to some richer-than-disquotational notion of truth in his theory of interpretation, the level at which he explains what meaning is. But at this level, too, truth talk turns out to be no more than the expressive convenience that functional deflationism takes it to be.

## V. TRUTH AND MEANING

For Davidson, a theory of meaning (narrow sense) takes the form of a Tarski-style truth definition. The instances of convention-T, the T-sentences, which emerge as theorems give the meaning of each individual sentence of the target language by stating that sentence's truth condition. In "Truth and Meaning," Davidson[15] advocates such a structure as the proper form for a theory of meaning because it avoids appeal to a nonextensional notion, such as 'means that'. Instead of requiring that a theory of meaning issue in theorems of the form '*s* means that *P*' for every sentence *S* of the target language, he suggests treating the position occupied by '*p*' extensionally, replacing 'means that' by 'if and only if' and supplying the sentence-identifying description '*s*' with some predicate proprietary to the theory of meaning. This yields theorems of the form:

(T) *S* is *T* if and only if *p*.

All we need now require of a theory of meaning for a language *L* is that, without appeal to any (further) semantical notions, it place enough restrictions on the predicate 'is *T*' to entail all sentences got from schema *T* when '*S*' is replaced by a structural description of a sentence of *L* and '*p*' by that sentence. Since the condition placed on satisfactory theories of meaning is essentially Tarski's convention-T, "it is clear that the sentences to which the predicate applies will be just the true sentences of *L*" (*ibid.*, p. 23).

This argument seems to imply that a Tarski-style truth definition could function as both a (partial) explanation of truth and an explanation of meaning (for a given language). In "Truth and Meaning," Davidson embraces this implication. Thus:

---

[15] In *Inquiries into Truth and Interpretation* (New York: Oxford, 1984), pp. 17-36.

> There is no need to suppress...the obvious connection between a definition of the kind Tarski has shown how to construct, and the concept of meaning. It is this: the definition works by giving necessary and sufficient conditions for the truth of every sentence, and to give truth conditions is a way of giving the meaning of a sentence. To know the semantic conception of truth for a language is to know what it is for a sentence—any sentence—to be true, and this amounts, in one good sense we can give to the phrase, to understanding the language.[16]

In later writings, however, Davidson argues that his position in "Truth and Meaning" embodies a confusion. His mistake, he says:

> ...was to think that we could both take a Tarski truth definition as telling us all we need to know about truth and use the definition to describe an actual language.[17]

We cannot both treat a Tarski-style truth definition as an empirical theory of meaning for a natural language and as telling us about truth, because we cannot assume an ability to recognize what T-sentences to count as correct. This was not a problem for Tarski: he simply assumed that, in a T-sentence, the sentence to the right of the biconditional was a correct translation of the sentence whose truth conditions it gives. But for Davidson, the main problem is to explain the empirical basis for a Tarski-like theory's being taken to give correct truth conditions for the sentences of some target language. That basis must consist, at least in part, in our ability to recognize that certain alien utterances are true, antecedently to our attaining a detailed understanding of what they mean. But this means that, in constructing a theory of meaning, we rely on a prior understanding of truth.

Davidson takes this argument to show that we face a fundamental choice: either we use the notion of truth to get at meaning or we explain truth in ways that rely on the notion of meaning. He takes the first option. He writes:

> ...while Tarski intended to analyse the concept of truth by appealing (in convention-T) to the concept of meaning (in the guise of sameness of meaning or translation), I have the reverse in mind. I considered truth to be the central primitive concept, and hoped, by detailing truth's structure, to get at meaning.[18]

Davidson sees himself as turning Tarski's theory on its head. Instances of convention-T are no longer "partial definitions" of truth, which is how many deflationists see them: they are empirical hypotheses about the truth conditions for sentences of a given speaker's language, hypotheses which require a prior understanding of the notion of truth. A theory of meaning illuminates meaning by revealing, in a systematic way, how the truth condition of a sentence is a function of the semantic properties of its parts: its logical structure and the reference of its terms.

---

[16] "Truth and Meaning," p. 24.

[17] "The Structure and Content of Truth," p. 286, note 20.

[18] "Introduction," *Inquiries into Truth and Interpretation*, pp. xiii-xx, here p. xiv.

Horwich accepts Davidson's appraisal of the situation. Thus, if:

> ...understanding a sentence...is a matter of appreciating what must be the case for the sentence to be true—knowing its *truth condition*—...it is not possible that this knowledge also helps to constitute our grasp of 'is true'. For in that case we would be faced with something like a single equation and two unknowns.... [K]nowledge of the truth condition of a sentence cannot simultaneously constitute *both* our knowledge of its meaning *and* our grasp of *truth* for the sentence.[19]

Horwich thinks that Davidson is right in the options that he sees and wrong only in the choice that he makes. We should explain meaning some other way—in terms of proper use. Then we can give a deflationary account of truth.

At first sight, this line of argument seems compelling. I think that Davidson's early "Truth and Meaning" position is much more defensible, however, than Davidson, in his later writings, allows. Are we really forced to choose between taking truth for granted so as to explain meaning or meaning for granted so as to explain truth? We must proceed carefully here. Even if Davidsonian theories of meaning require some prior grasp of truth, it does not follow that the function of truth in such theories is to be explanatory of meaning. Accordingly, it does not follow that such theories could not tell us something about meaning and something (though not everything) about truth.

A Davidsonian theory of meaning, for a particular speaker, is a device for *giving the meaning* of every sentence in that speaker's idiolect. If we like, we can say that such a theory explains what that speaker's sentences mean. But to explain what particular sentences mean is not to explain what meaning is. A fortiori, it is not to explain what meaning is by appeal to truth. In this respect, Davidsonian theories of meaning are just like Quinean translation manuals. Translation manuals give translations; they do not explain what translation is or what makes one sentence a good translation of another. The illusion that Davidsonian theories of meaning are explanatory is fostered by loose talk about "explaining meaning in terms of truth conditions" or "identifying a sentence's meaning with its truth condition." For Davidson, there are no such things as meanings, hence no question of identifying them with anything. Davidsonian meaning theories are more accurately described as *specifying* meanings or, better still, as *interpreting* the sentences of some target language.

---

[19] *Truth*, p. 71.

In specifying the meanings of sentences in another speaker's idiolect, we associate his sentences with sentences of our own. In doing so, we make use of the truth predicate, which is what lends color to the idea that Davidson explains meaning in terms of truth conditions. But the use of 'true' in a Davidsonian meaning specification for a particular speaker is expressive, not explanatory. He needs 'true' only because he eschews 'means that' in favor of the material biconditional. Making use of 'means that', we could specify meanings using such sentences as:

> '"Le Président de France est fou" means that the President of France is mad'.

But:

> '"Le Président de France est fou" if and only if the President of France is mad'.

is not even grammatically well formed. To replace 'means that' with a sentential connective, we need a sentence on the left side. An ideal candidate would be a prosentence referring anaphorically to the target-language sentence whose meaning we want to specify. This is precisely what 'true' allows us to form. The appeal to truth in a Davidsonian theory of meaning is not explanatory.

## VI. REFERENCE RELATIONS AS THEORETICAL ENTITIES

For Quine, correct translations are all and only those which satisfy the methodological constraints enshrined in the theory of translation: there is no further fact of the matter, in respect of meaning, for translation manuals to be right or wrong about. With respect to meaning, Quine is a thoroughgoing instrumentalist. But so is Davidson. Substituting 'interpretation' for 'translation', what we just said about Quine goes for Davidson, too. Davidson is not a verificationist in the theory of meaning (narrow sense). That is, individual theories of meaning give meanings by way of truth conditions (in the innocuous sense indicated in the previous section): they make no mention of verification or assertibility conditions. But he is surely some kind of a verificationist about the subject matter of such theories. This follows from his Quinean antirealism with respect to meanings.

This feature of Davidson's thought about meaning is obscured by an analogy to which he appeals to clarify the way in which truth is his real explanatory primitive. This analogy also contributes powerfully to the illusion that Davidsonian theories of meaning are themselves explanatory. The analogy in question is between reference relations and the unobservables postulated by theoretical science.

Davidson invokes this analogy in the course of responding to the suggestion that, since Tarski shows how to define truth in terms of reference, we could take the notion of reference as primitive. Davidson rejects this thought on the grounds that the basic evidence for theories of meaning is given by sentences held true in observable circumstances. We therefore decide what the basic terms of another person's language refer to by determining what assignment of referents make his sentence held true come out to be actually true. This means that in constructing a theory of meaning, even though it takes the form of a Tarskian theory of truth, we make use of a notion of truth that is not defined by the theory. The crucial point here, as Davidson points out, is not "to confuse the order of explanation that is appropriate once the theory is in place with the explanation of why the theory is correct." Thus:

> The theory is correct because it yields correct T-sentences; its correctness is tested against our grasp of the concept of truth as applied to sentences. Since T-sentences say nothing about reference, satisfaction, or expressions that are not sentences, the test of the correctness of the theory is independent of intuitions concerning these concepts. Once we have the theory, though, we can explain the truth of sentences on the basis of their structure and the semantic properties of the parts. The analogy with theories in science is complete: in order to organize and explain what we directly observe, we posit unobserved or indirectly observed objects and forces; the theory is tested by what is directly observed.[20]

This makes brilliantly clear how, for Davidson, reference in one way determines truth while at the same time, in another way, it is determined by it. A Davidsonian theory of meaning, while formally compositional, is methodologically holistic. As he says,

> If sentences depend for their meaning on their structure, and we understand the meaning of each item in the structure only as an abstraction from the totality of sentences in which it features, then we can give the meaning of any sentence (or word) only by giving the meaning of every sentence (and word) in the language.[21]

---

[20] "The Structure and Content of Truth," p 300.

[21] "Truth and Meaning," p. 22.

But because Davidson treats the reference of a term as "an abstraction from the totality of sentences in which it features," reference has no life of its own, independent of the truth of sentences. The analogy with the theoretical entities postulated by physicists is thus misleading. The analogy implies that the dependence of reference on linguistic behavior as a whole is merely epistemological, and that is not Davidson's view. Reference relations are not independently existing objects that explain observable features of a person's linguistic behavior. Rather, they are *constituted* by their role in making a person's utterances, or most of them, come out true. They have no further properties for us to speculate about. In this respect, they are quite unlike electrons.

## VII TRUTH AND INTERPRETATION

We have seen that Davidsonian theories of meaning (narrow sense) are specifications of meaning. As such, they make essential use of the expressive power of the truth predicate. But they do not explain what meaning is: that is the task of the theory of interpretation, which articulates the constraints that determine when such meaning specifications can be regarded as correct. The question, therefore, is: Must Davidson take truth as an explanatory primitive in the theory of interpretation?

Davidson says that, if we ask "What is missing, as an account of truth, in a theory of truth in Tarski's style?" the answer is "the connection with users of language." This is just the question of how we decide "whether a theory of truth is true of a given language." But:

> since all of us do understand some speakers of some languages, all of us must have adequate evidence for attributing truth conditions to the utterances of some speakers; all of us have, therefore, a competent grasp of the concept of truth as applied to the speech behavior of others.[22]

But what does this grasp consist in? If we know how to interpret, we can extend our purely disquotational notion of truth to alien sentences, which would appear to give us just what Davidson demands. So the question remains: Why not give truth to the deflationists and let interpretation stand on its own feet? The only objection to so doing would have to be that the canons of interpretation themselves make use of a rich notion of truth. I do not believe that they do.

Davidson's views about interpretation can be seen as consequences of his holism. Thus, the first constraint on interpretation is that we must find semantic structure—logical form—in the sentences of our target language. This is, of course, an immediate consequence of holism, which treats the meaning of individual words as an abstraction from their contribution to the meanings of all the sentences in a person's language. But to see words as making a systematic contribution, we must be able to identify the logical form of the sentences in which they occur, and to see sentences as standing in logical relations to one another. According to Davidson, this is how interpretation differs from translation.

---

[22] "The Structure and Content of Truth," p. 301.

Davidson refers to finding logical form in someone's utterances as locating "the kind of pattern truth must make" and he thinks that Tarski's contribution is to have shown just what that pattern must be. But what Tarski showed is how to define truth in terms of reference for languages formalized in first-order logic. So in practice, finding the pattern that truth must make means using standard first-order logic as the basis for a theory of the logical forms of sentences of natural languages. This will involve showing how seemingly recalcitrant sentences—such as those involving adverbial modification—can be accommodated within this general structure. A commitment to finding this particular kind of logical structure in the sentences of natural languages has, however, no essential connection with a nondeflationist view of truth. We *could* say that adopting a particular logic amounts to adopting a particular—and so richer-than-disquotational—concept of truth. But we do not have to; and nothing in Davidson's view of meaning, which requires only that we find in natural languages the structure of first-order logic, mandates this further step. Davidson's theory of interpretation privileges first-order logic. It does not, however, require that we build our logical preferences into our concept of truth.

The second constraint is charity: we must take the beliefs of anyone we are trying to interpret to be, on the whole, true. What notion of truth is involved here?

In his early writings, Davidson sees two dimensions to charity. One dimension—consistency—follows from the need to find logical structure. Logical relations between sentences are our clues to logical form. But such clues would be useless in the face of massive illogicality. More deeply, in the face of such illogicality, there would be no pattern of logical interrelationships, hence no meaning, nothing to discover. But this imports no richer notion of truth than is already involved in the need to read a logic into the language of our interlocutor.

What about the meaning of nonlogical terms? Given holism, the meaning of a person's words has to be abstracted from sentences he holds true, that is, from his beliefs. But we cannot know what his beliefs are unless we can understand his words. Meaning and belief are thus interdependent, another equation in two unknowns. The way to solve the problem of the interdependence of belief and meaning is by:

> ...holding belief constant while solving for meaning. This is accomplished by assigning truth conditions to alien sentences that make native speakers right when plausibly possible, according, of course, to our own view of what is right. What justifies this procedure is the fact that agreement and disagreement alike are intelligible only against a background of massive agreement.[33]

This is the second dimension of charity: we have to use our own beliefs as the basis for interpreting the beliefs of others. Clearly, however, the key notion here is agreement, not truth. Indeed, Davidson himself often explains charity in just these terms, as when he says that the crucial point is that "the correct interpretation of an agent by another cannot intelligibly admit certain kinds and degrees of difference between interpreter and interpreter with respect to belief."[34] This says everything that needs to be said, without truth getting so much as a mention.

To be sure, Davidson does also formulate his key methodological maxim in ways that make use of the notion of truth. Thus: "Making sense of the utterances...of others...requires us to find a great deal of reason and truth in them."[35] It is clear, however, that the concept of truth, as it occurs here, bears no explanatory load. Charity requires us to attribute to the alien the maximal possible number of beliefs that we ourselves hold true. So our maxim is: for the most part, do not attribute to the alien the belief that *p* unless *p*, the belief that *q* unless *q*, and so on. This maxim employs 'true' as the generalizing device that deflationists say it is.

Davidson thinks that deflationists sever all links between meaning and truth. But to the extent that such a link is enforced by charity, this is simply not so. Nothing prevents a deflationist about truth from holding that interpretation requires attributing to other people commitment to lots of truths. And this is all charity involves. To find a great deal of truth in the utterances of another is just to find among those utterances lots of truths. Since, inevitably, these will be truths "by our own lights" or "according to our own standards," being charitable involves only what Field calls "mapping lots of alien sentences onto disquotationally true sentences that we ourselves understand" (*op. cit.*, p. 408). Charity can rejoice in truth without being puffed up.

---

[33] "Radical Interpretation," *Inquiries into Truth and Interpretation*, pp. 125-39, here p. 137

[34] "The Structure and Content of Truth," p. 319.

[35] "Belief and the Basis of Meaning," in *Inquiries into Truth and Interpretation*, pp. 141-54, here p. 153

There is one final aspect of Davidson's views to consider. To fix the interpretation of nonlogical terms, charity as consistency-plus-truth-maximization is not enough: we need a constraint on interpretation that connects language with the world. Here, Davidson follows Quine. The fundamental evidence to which interpretation is responsive is given by occasion sentences that are responsive to readily detectable events in our surroundings: the passing by of a rabbit and so on. As Davidson himself remarks, the distinction between "sentences whose causes to assent come and go and those a speaker clings to through change" is all that "offers the possibility of interpreting the words and sentences beyond the logical."[16] So we have a third dimension of charity: attributing true beliefs in certain methodologically basic cases.

Is there a rich notion of truth here? Again, no. As before, the relevant interpretive maxim involves a generalizing rather than an explanatory use of 'true'. With respect to observation sentences, the rule is: interpret 'Gavagai' as 'Lo a rabbit' if there is a rabbit passing by; or in general, make the interlocutor's observation sentences come out true. Notice, incidentally, that, in the case of observation sentences, nothing distinguishes Davidson's theory of interpretation from an approach to meaning in terms of assertibility conditions. In the case of observation sentences, assertibility conditions include truth conditions; and to state such assertibility conditions, we only need 'true' as a generalizing device.

Let us return to Horwich's suggestion that deflationists, since they cannot explain meaning in terms of truth, should explain it in terms of use. If I am right that the role of truth in Davidsonian theories of meaning is expressive rather than explanatory, and if it turns out that the notion of truth plays no essential explanatory role in the theory of interpretation either, Davidson may be closer to Horwich than either thinks. Or I should say "used to think," for in his latest writings Horwich[17] has come round to the view that "Davidson's Principle of Charity arguably boils down to the use theory of meaning" (*ibid.*, p. 72). I think Horwich is right: it does.

The theorems of a Davidsonian theory of meaning are material biconditionals. Since such biconditionals will be true provided that they pair true sentences, it is natural to wonder how they can be regarded even as specifying meanings. Given a material reading of the biconditional:

---

[16] "A Coherence Theory of Truth and Knowledge," in Lepore, ed., *Truth and Interpretation*, pp 307-19.

[17] *Meaning* (New York: Oxford, 1998), p 72.

(S) '"Snow is white" is true if and only if grass is green'

is true, but 'Snow is white' does not mean 'Grass is green'. But this is why Davidson embraces holism. (S) is not absurd in itself. It is just that we would be entitled to come up with (S), in an attempt to understand a fellow speaker, only if we could construe lots of that speaker's other utterances, subject to the multiple constraints of charity: in particular, those relating to his observational uses of 'snow', 'white', and 'green'. Sentences giving "truth conditions" can be regarded as specifying meanings only in the context of a scheme for interpreting an entire language. Seen in this light, Davidson's view of meaning does not stand in a simple relation of opposition to what Horwich conceives of as a use theory. As Horwich puts the point: "Implicit in...the Principle of Charity...are the following ideas: (1) that what a person holds *true* is determined by various factors, including the observable circumstances, background beliefs, inferential propensities, and basic regularities of use of words—regularities that specify, as a function of the other factors, which sentences are held *true*; and (2) that a good assignment of truth conditions should optimize agreement in the sense of preserving these basic regularities."[illegible] And since, by Davidson's lights, the reference of a speaker's terms is no more than an abstraction from the totality of the sentences he holds true, reference itself is a matter of contribution to or function within a belief system, as manifested in patterns of assent and dissent, special attention being paid to occasion sentences. In a word, reference is a matter of use.

## VIII. CONCLUSION

What do the T-sentences that are the theorems of a theory of meaning for a particular language or speaker tell us about: truth or meaning? To this question the final answer is "both."

They tell us something about truth: they fix the extension of 'true' as applied to alien sentences, taking us from pure to extended disquotational truth. They do not, however, "explain what truth is." They presuppose a grasp of the function and mechanism of truth in

---

[illegible] *Ibid.* Despite this take on charity, Horwich continues to think of the problem of ruling out aberrant T-sentences of the '"Snow is white" is true if and only if grass is green' variety as a possibly insuperable problem for Davidson Perhaps he reads too much into the claim that, for Davidson, an utterance's truth condition "constitutes its meaning"—see, for example, *ibid.*, p. 168, footnote 11.

our home language, exploiting without explaining the expressive function that truth talk serves.

Davidson's account of when such theories are to be counted correct makes no appeal to notions like "objective interpersonal synonymy." Our "meaning theory" for a given speaker undergoes constant modification and so is really a series of "passing theories."[29] Davidson therefore respects Field's constraint that a deflationist should treat "'good translation' as a highly context-sensitive and interest-relative notion."[30] For Davidson, 'correct' here means acceptable. It does not and should not mean 'uniquely acceptable'. Davidsonian theories of meaning offer a way of moving from pure to extended disquotational truth which meets Field's standards for being (theoretically) deflationary.

The theorems of a Davidsonian theory of meaning also tell us something about meaning: they specify the meanings of the sentences of the target language by pairing them systematically with sentences antecedently understood. But they do not explain what meaning is: that is accomplished by the theory of interpretation, which explains what makes such a systematic pairing acceptable.

There is no circularity here: meaning explaining truth and truth explaining meaning. Rather, Davidsonian theories of meaning derive their power to tell us something about meaning and something about truth from the theory of interpretation, in which neither meaning nor truth figures as a primitive notion. Davidson thinks that he explains meaning in terms of truth. But what he really does is illuminate both meaning and truth by appeal to interpretation, an activity that can itself be understood without extensive theoretical resources. This is why he can afford to be a deflationist. For someone like me, attracted to both a Davidsonian approach to meaning and a deflationary approach to truth, this is a welcome result.

MICHAEL WILLIAMS

Northwestern University

---

[29] See Davidson, "A Nice Derangement of Epitaphs," in LePore, ed., *Truth and Interpretation*, pp. 125-39.

[30] Field, p. 408.

# KNOWING FAILABLY*

Almost all epistemologists are nonskeptics: they accept that people have knowledge. Almost all epistemologists say, too, that most—maybe all—human knowledge is fallible. But what is it for an instance of knowing to be fallible? The usual answer will be something like this:

> FalK: one's knowledge that *p* is fallible, if and only if (1) one knows that *p* but (2) one's justification for one's belief that *p* is compatible with one's being mistaken in believing that *p*.[1]

Yet this is not a wholly satisfactory characterization. No belief in a necessary truth can be mistaken—because no necessary truth can be false. So FalK entails that no such belief could be fallible knowledge; and that seems like a mistaken result itself. One's justification for a belief in a necessary truth could be less than perfect, and this should make the belief fallible knowledge at best—even though this justification is incompatible with the belief being mistaken. Should we therefore broaden FalK, so that a piece of knowledge that *p* can be fallible because of something *other* than *p*'s possible falsity (in some pertinent sense of 'possible')?

In my experience, epistemologists will be reluctant to do that. Fallibility (they will say) is definitionally tied to a possibility of being mistaken. The problem with that reluctance to seek a broader analysis of fallible knowledge, however, is that it can prevent our noticing a more general kind of epistemic failing—in fact, the kind of failing that underlies the traditional concept of knowing fallibly. Were it not for that reluctance, I would be happy to extend the term 'fallible knowledge' to this more general kind of failing. Given that reluctance, however, I shall call this more general phenomenon epistemic *failability*—a way of knowing which generates what I call *failable knowledge*. And it will be the subject of this paper.[2]

---

* I am grateful to Scott Campbell, Richard Feldman, Tim Oakley, Phillip Staines, and Konrad Talmont-Kaminski for their helpful comments on an earlier version of this paper

[1] See, for example, Stewart Cohen, "How To Be a Fallibilist," in James E. Tomberlin, ed., *Philosophical Perspectives*, Volume 2 (Atascadero, CA: Ridgeview, 1988), pp. 91-123, here p. 91.

[2] Philosophers use the term 'fallibilism' in two ways—to denote fallible knowledge, or to refer to human cognitive fallibility in general. See, for instance, Susan Haack and L. S. Carrier in, respectively, "Fallibilism and Necessity," *Synthese*, XLI, 1 (May 1979): 37-63, and "How to Define a Nonskeptical Fallibilism," *Philosophia*, XXII, 3-4 (December 1993). 361-72. I shall eschew the term 'fallibilism' and its cognates, as used in the second of these two ways. Instead, I shall be generalizing the *first* of those two ways.

0022-362X/99/9611/565–87

## I. FAILABLE KNOWING IN GENERAL

(A) The underlying motivation for the usual epistemological analysis of what it is to know fallibly is clear. The motivation is the fact that possible falsity is a possible lack of knowledge. Because knowledge is of only what is true, a given *p*'s being false entails that no one knows that *p*. So, any belief that *p* (for that given *p*) could have *failed* to be knowledge—even when, in fact, it *is* knowledge.

With that observation, though, we are already on our way to a broader, hence more widely applicable, conception of knowing-in-a-less-than-ideal-way than is provided by the usual conception of fallible knowing. Instead of restricting ourselves to an analysis of fallible knowing, we should focus on this more general idea:

> FailK1: one's knowledge that *p* is failable, if and only if (1) one knows that *p* but (2) one might have failed to do so.

Knowing *failably* entails possibly not knowing. And knowing *fallibly* entails knowing failably: if one's belief that *p* is knowledge yet it could be mistaken, then it is knowledge which could fail to be knowledge, and hence it is knowledge which might not have been knowledge. I shall accept FailK1 as a minimal—formal, generic, schematic—condition to be satisfied by any more specific account of failable knowing.

(B) In order to find a more specific instantiation of FailK1, we need only describe the various routes to a possible lack of knowledge (the possible lack referred to by FailK1 (2)). Possible falsity is one of them, but there are also other ways possibly to lack a piece of knowledge. Let us assume that knowing that *p* involves having at least a well-justified true belief that *p*. That being so, one lacks knowledge that *p* if *p* is false, or if one does not believe that *p*, or if one lacks good justification for the belief that *p*. Consequently, there are at least the following three examples of ways in which one might not have known that *p*.

(i) It could be the case that one's evidence for *p* is good, but not so good as to be incompatible with *p*'s being false. In that event, *p* might not have been true, even if it is true and even if the evidence for its truth is itself true. This is a way for knowing to be failable due to one's relying on nondeductive justification—that is, due to its being fallible.

(ii) It could be the case that one believes that *p*, but not so confidently or firmly that one might not have decided not to have that belief, even though the belief is true and even if in fact the belief is produced in response to one's good evidence for its truth. This is a way for knowing to be failable due to one's belief being present in too fragile a way, let us say.

(iii) It could be the case that even though one has good evidence for one's true belief that $p$ and even though one does base one's belief that $p$ upon that evidence, one might have had that true belief anyway, even without that evidence for it. This is a way for knowing to be failable due to one's belief being present in too dogmatic a way, let us say.

Each of (i)-(iii) describes a *minimal* possible way to fail to know that $p$. Each is minimal, because each describes one's lacking just one of the three features described so far as being required for knowing. In (i), truth is absent ($p$ is false), while the belief that $p$ and good justification for that belief are present. In (ii), the belief that $p$ is absent, while $p$ is true and one has good justification for that belief. And in (iii), one lacks good justification for the belief that $p$, while having a true belief that $p$.

Of course, one's knowing on a given occasion would also be absent if one lacked even more than just one of those three features (and of any other features that knowledge requires, if there are any); subsection (D) will discuss that kind of situation. But our first need here is for a *core* conception of knowing failably. That is why I have outlined some minimal possible ways to fail to know that $p$. What will prove in section II to be a useful heuristic with which to sum up the core analysis I am about to present, then, is this: to know failably that $p$ is to know that $p$ even while coming at least a little close to not knowing that $p$. More formally:

> FailK2: a person $x$ knows failably that $p$, if and only if (1) $x$ knows that $p$, and (2) there is an accessible possible world where (i) $p$ is false (but $x$ believes that $p$, with the same good evidence for $p$ as he has here), or (ii) $x$ fails to believe that $p$ (even though $p$ is true and he has the same good evidence for $p$ as he has here), or (iii) $x$ fails to have the same good evidence for $p$ as he has here (but he still believes that $p$, and $p$ is true).[3] (And an accessible possible world, as regards $x$'s

---

[3] Should clause (iii) be the following more sweeping condition instead?

$x$ fails to have good evidence for $p$ (but he still believes that $p$, and $p$ is true)

I do not believe that it should be. I am discussing whether a particular piece of knowledge that $p$ is failable. From FailK1, it is failable if it could be absent (while the person is still present); and a minimal way for it to be absent is for there to be an accessible possible world where the person's true belief that $p$ is supported by different good evidence. For that is a world where at best the person has a different case of knowing that $p$. From FailK1, to know infailably on a given occasion is for that knowing—characterized by whatever features make it the instance of knowing it is—to be infallible. And the features that make it the instance of knowing it is include at least the particular belief it is, and that belief's being true, along with the particular good evidence for it. Hence, if there is to be identity of knowing across worlds, then at least the same belief and truth and good evidence are required in each such world. (But if you reject transworld identity of knowing, you might replace clause (iii) with the suggested alternative. That would affect some subsequent parts of this paper in detail but not principle.)

> knowing that $p$ in this world, is a world which contains at least two members of this set: {$p$'s being true, $x$'s believing that $p$, $x$'s having the same good evidence for $p$ as he has in this world}.)[4]

FailK2 entails that for $x$ to know infailably that $p$ is for there to be not even a *possibility* of $x$'s lacking that knowledge. And FailK2 describes three minimal ways to instantiate that possibility. I shall refer to worlds containing those instantiations as being *epistemic-failure* worlds for $x$ in relation to $p$.[5] Failable knowledge is knowledge for which there is a possible epistemic-failure world.

(C) Is my account of infailably knowing really an account of necessarily knowing? After all, the account begins with FailK1, which says that infailable knowing is knowing which is not such that it might not have existed. That is (it seems), one's infailably knowing that $p$ is one's necessarily knowing that $p$—where one's necessarily knowing that $p$ is either (i) in every possible world, one's knowing that $p$, or (ii) in every possible world where one exists, one's knowing that $p$.

In fact, FailK1 does not have that commitment. FailK1 has led to FailK2—an account which describes infailable knowledge's infailability component in this way:

> In each world where one exists and where one has two of the elements of knowing that $p$, one also has the third element.[6]

Necessarily, to be that close to knowing that $p$ *is* to know that $p$; at least it is, when the knowing is infailable. That is what FailK2 implies—clearly, a weaker commitment than is involved in necessarily knowing.[7]

---

[4] Any epistemic-failure world for $x$ in relation to $p$ is thus an accessible world as regards $x$'s knowing that $p$. (Hence, when I talk below of epistemic-failure worlds, I shall not need to make explicit that they are accessible worlds. And because any world that satisfies one of (i)-(iii) in FailK2(2) is an epistemic-failure world for $x$ in relation to $p$, it is also an accessible world, as regards $x$'s knowing that $p$. So, strictly, there was no need to say, as FailK2(2) does, that it is referring to accessible worlds.) But the converse implication fails, because if $x$ knows *infailably* that $p$ then any world containing at least two members of the set {$p$'s being true, $x$'s believing that $p$, $x$'s having the same good evidence for $p$ as he has here} also contains the third member of the set. That is, if $x$ knows infailably that $p$, then none of the pertinent accessible worlds is an epistemic-*failure* world for $x$ in relation to $p$.

[5] Implicit in this, too, is a relativization to a particular time—a world being an epistemic-failure world for $x$ in relation to $p$ *at time* $t$. But, for simplicity and because no unclarity will result, I ignore this extra relativization in what follows.

[6] "Does knowing not have more than three elements—truth, belief, good justification, and something else? Did Gettier not teach us that?" If section II below is correct, we do not have to agree that Gettier did that. (And even if section II is wrong, the present discussion can be expanded to accommodate whatever else is involved in knowing.)

The usual analysis—FalK—of fallible knowledge that $p$ describes it as being knowledge that $p$ which could be absent via the truth that $p$ being absent, even when the belief that $p$ and the good justification for that belief are present. In my terms: it describes fallible knowledge that $p$ as knowledge that $p$ which is compatible with (i) being that close to knowing that $p$, yet (ii) still not knowing that $p$. I am generalizing that conception of fallible knowledge to obtain the concept of failable knowledge—so that failable knowledge that $p$ is knowledge that $p$ which could be absent via any one of the truth that $p$, the belief that $p$, and the good justification for that belief being absent, even when the other two are present. Can a person be that close, in any of those respects, to knowing that $p$—yet without knowing that $p$? The distinctive feature of one's infailably knowing that $p$ is its being impossible for one not to know that $p$ once one is so close to having the knowledge that $p$.

(D) Although infallibility is absolute, fallibility is not. It can be more or less. We talk of a person's being more, or his being less, fallible, at any rate: he is more, or he is less, likely to be mistaken, even given his evidence. The same way of talking is appropriate for epistemic failability. A person can be more, or he can be less, epistemically failable: he is more, or he is less, likely to fail to know various facts. And can that qualitative or comparative dimension of epistemic failability be applied to a particular instance of knowing? Can one person know that $p$ *more* failably than another person does? Can one know that $p$ more failably today than yesterday?

That certainly can occur, at least in principle. If (as FailK2 tells us) $x$'s knowing that $p$ being infailable involves there being no epistemic-failure worlds for $x$ in relation to $p$, then $x$'s knowing that $p$ being more, or its being less, failable involves (respectively) there being more, or there being fewer, and there being closer, or there being more distant, epistemic-failure worlds for $x$ in relation to $p$. More fully:

---

[7] Haack presents dogmatism—the contradictory of what she calls fallibilism—as saying that "there are some propositions which we can't fail to believe *and which* are such that it follows from our believing them that they are true" (*op. cit.*, p. 52). Adapting that to our present concerns gives us this:

> $x$'s knowing infallibly that $p$ includes (i) $x$'s believing that $p$ in every world where $x$ exists, and (ii) $p$'s being true in every world where $x$ believes that $p$

That is, in every world where one exists, one has a true belief that $p$—this being so, for any $p$ which one infallibly knows. My account of infailable knowing is less demanding than that account of infallibly knowing. I require that, if one is infailably to know that $p$, then each world where one exists *and* which includes at least two members of this set—{$p$'s being true, one's believing that $p$, one's having good evidence for that belief}—also includes the third member.

> FailK3: a person $x$ knows more, or $x$ knows less, failably that $p$, if and only if (1) $x$ knows that $p$, and, respectively, (2) there are more, or there are fewer, and there are closer, or there are more distant, possible worlds where (i) $p$ is false (but $x$ believes that $p$, with the same good evidence for $p$ as he has here), or (ii) $x$ fails to believe that $p$ (even though $p$ is true and he has the same good evidence for $p$ as he has here), or (iii) $x$ fails to have the same good evidence for $p$ as he has here (but he still believes that $p$, and $p$ is true).

Exactly (or even approximately) how failable a particular piece of knowledge is will be a function of the pertinent epistemic-failure worlds, the worlds that satisfy clause (2) of FailK3. Correlatively, whether one piece of knowledge is more failable than another is also a function of those worlds. If $x$ knows that $p$ and knows that $q$, but there are more close epistemic-failure worlds for $x$ in relation to $p$ than in relation to $q$, then—other things being equal—$x$'s knowledge that $p$ is more failable than his knowledge that $q$.[8]

For example, right now I know failably that I am typing. I also know failably that 13 x 14 = 182. As far as I can tell, though, the former piece of knowledge is more failable than the latter one. In rather rough outline, here is why that is so.

(a) There are possible worlds where I am not typing even though I believe that I am and even though I have the same sensory evidence as I have here for my typing. (In those worlds, I am hallucinating, say.) There are worlds where I am typing and where I believe that I am, but where my evidence is not as it is here. (In those worlds, I am sensorily deprived, yet I remember intending to be typing now, say.) There are worlds where I am typing and where I have the same sensory evidence as here, but where I fail to believe that I am typing. (In those worlds, I distrust that evidence, say.) These are thus three kinds of epistemic-failure world for me in relation to the fact of my typing.

(b) There are possible worlds where I have the same mathematical evidence as here for the fact that 13 x 14 = 182, but where I fail to believe that 13 x 14 = 182. (In those worlds, my caution about such calculations prevents me from having that belief, say.) There are worlds where I have the belief without that evidence. (In those worlds, I remember being told that $13 \times 14 = 182$, say.) But there are no worlds where $13 \times 14 \neq 182$.

---

[8] Could $x$ also know that $p$ more failably on one occasion than on another? If so, would his more failable knowledge that $p$ be less good, purely qua knowledge that $p$, than his less failable knowledge that $p$? The idea that propositional knowledge can admit of such qualitative gradations is rarely countenanced by epistemologists. I have accepted it elsewhere—"The Sceptic Is Absolutely Mistaken (As Is Dretske)," *Philosophical Papers*, XXVII, 1 (March 1998): 29-43. David Lewis has also endorsed that idea in "Elusive Knowledge," *Australasian Journal of Philosophy*, LXXIV, 4 (December 1996): 549-67, here pp. 562-63.

All else being equal, then, there are more epistemic-failure worlds for me in relation to the fact of my typing than for me in relation to the fact that $13 \times 14 = 182$. I have distinguished between three kinds of epistemic-failure world. We could call them (in turn) *truth*-failure worlds, *belief*-failure worlds, and *justification*-failure worlds. For me in relation to the fact of my typing, there are truth-failure worlds, belief-failure worlds, and justification-failure worlds. For me in relation to the fact that $13 \times 14 = 182$, there are only belief-failure worlds and justification-failure worlds—no truth-failure worlds. Therefore, all else being equal, my knowledge of my typing is more failable than my knowledge that $13 \times 14 = 182$.[9]

## II FAILABLE KNOWING WITHIN GETTIER CASES

(A) If failable knowing can be more or less failable, then in principle there can be instances of failable knowing which are very failable. These would be instances where the person is not just a little close to failing to know; instead, he is very close to lacking the knowledge. There are simple ways for that to occur (for example, I look up at just the right time to see, and hence know, who was running away from the robbery). But there are more surprising ways, too, in which knowing can be like that. The phenomenon of Gettier cases[10] supplies us with one such example.

All epistemologists know (albeit failably) that such cases are regarded by all epistemologists as ones where there is a lack of knowledge.[11] Regardless of that uniformity, I wish to outline an alternative way of interpreting the cases, one according to which there is knowledge within the cases—failable knowledge, indeed very failable knowledge.

Indisputably, each Gettier case contains some significant luck on the part of the case's epistemic subject. His evidence has led him to a true belief that *p*; and the evidence is, in itself, good evidence for that belief. Nevertheless, each case also contains what we may call a

---

[9] And *is* all else equal? It seems to me that I am more likely not to be typing, even when I think that I am doing so, than I am to suspend my simple mathematical beliefs when I engage in apparently normal calculations. In other words, the truth-failure worlds for me in relation to the fact of my typing are closer possible worlds than are the belief-failure worlds for me in relation to the fact that $13 \times 14 = 182$. If so, then on this score, *too*, my knowledge of my typing is more failable than my knowledge that $13 \times 14 = 182$.

[10] Such cases either are, or trace their ancestry to, those found in Edmund Gettier's modern classic, "Is Justified True Belief Knowledge?" *Analysis*, XXIII, 6 (June 1963): 121-23. I shall assume some familiarity with such cases on the part of the readers.

[11] The conventional view of Gettier cases is that each contains an epistemic subject who has a belief which is true and well justified but not knowledge.

*strange occurrence.* And the presence of the strange occurrence is why the epistemic subject is somewhat lucky to have his well-justified true belief. That luck is manifested in one of the two following ways.

*(1) Helpful Gettier cases.* Some Gettier cases contain a strange occurrence which functions helpfully. If it had been absent,[12] then (all else remaining equal) the case's epistemic subject would not have had his well-justified true belief. He would have lacked the truth or the belief or the good evidence. He would not have been even as close as he is to having knowledge. For instance, Roderick Chisholm's[13] Sheep Case is like this. If there had not been a sheep in the field, hidden from his view, then (all else being equal) the epistemic subject's belief that there is a sheep in the field would not have been true. (It would have been supported by the same good but misleading sensory evidence as at present, seemingly of a sheep, really of a disguised dog.) Edmund Gettier's own cases are most likely helpful Gettier cases, too (*op. cit.*). In the first of those cases, for instance, if Smith had not been the person who will get the job, then (all else being equal) it is probable that his belief that the person who will get the job has ten coins in his pocket would have been false. (Because all else is equal, Jones still does not get the job; hence, someone other than Smith and Jones does. And it is probable that whoever does get the job will not have ten coins in his pocket.) Those are just two examples, but they are examples of a clear pattern: in each helpful Gettier case, it is true that, if the case's strange occurrence was absent, then—all else being equal—the case's epistemic subject would be in a close possible world where the truth that *p* or his belief that *p* or his good justification for *p* would be absent. In short: he would be in a pertinent and close epistemic-failure world for himself in relation to *p*. With the strange occurrence being present, though, the epistemic subject is not in an epistemic-failure world for himself in relation to *p*. (The strange occurrence plays a crucial part in bringing about his well-justified true belief.)

*(2) Dangerous Gettier cases.* Each remaining Gettier case contains a strange occurrence which functions as a hidden danger, an unseen threat, to the epistemic subject's having his well-justified true belief. If the strange occurrence had been absent, then (all else remaining

---

[12] Sometimes the strange occurrence is conjunctive, involving a few circumstances occurring. Always, though, at least part of the conjunction is vital to the case, in the way I am about to describe.

[13] *Theory of Knowledge* (Englewood Cliffs, NJ: Prentice-Hall, 1989, third edition), p. 93.

equal) the epistemic subject would not have been in any real danger of not having his well-justified true belief. He would have had his well-justified true belief (as he does already)—but in a wholly normal way (as he does not do already, within the Gettier case). He would have been as close as that can take him to having knowledge in a normal way. For instance, Carl Ginet's Fake Barns Case[14] is like this. Henry is looking at a real barn, just as he would be doing if there were no fake barns nearby. Of course, there *are* fake barns nearby: that is the case's strange occurrence. But now imagine that occurrence's absence. If there were no fake barns nearby, then (all else being equal) there would be no imminent danger of Henry's being deceived by a fake barn into believing that he is seeing a barn. Because all else is equal, he would still have that belief; and it would still be true, since it is a response to a real barn in front of him. The belief would still be well justified, too, by his standard sensory evidence for it. Accordingly, with the absence of the fake barns and with all else being equal, Henry would have a belief which is true and well justified in a normal way. Nevertheless, in this case—as in each dangerous Gettier case—the case's strange occurrence threatens to interfere: there are close possible worlds where it does interfere, these being worlds where the truth that $p$ or the epistemic subject's belief that $p$ or his good justification for $p$ are absent. For example, there is a close possible world where Henry *is* deceived by a fake barn. That is, there is a pertinent and close epistemic-failure world for Henry in relation to his seeing a barn.[15]

In each Gettier case, therefore, the strange occurrence brings it about that the epistemic subject almost fails to have his well-justified true belief. In a helpful Gettier case, he has a well-justified true belief, in part due to the occurrence of the strange occurrence—but, by the same token (the strangeness of the strange occurrence), he almost fails to have the well-justified true belief. In a dangerous Gettier case, on the other hand, the epistemic subject has a well-justified true belief *in spite* of the strange occurrence. In a helpful case, it would have been so easy for the strange occurrence not to have been present—and hence (all else being equal) for the well-justified true belief to have been absent. In a dangerous Gettier case, it would have been so easy for the strange occurrence not to have allowed the well-justified true belief to be present. So, with each kind of Gettier case, there are close possible worlds where at least one member of the crucial triumvirate—the truth that *p*, the epistemic subject's belief that *p*, and his good justification for that belief—is absent. In other words, for each kind of Gettier case there are pertinent and close epistemic-failure worlds for the case's epistemic subject in relation to *p* (the propositional object of his well-justified true belief).

---

[14] The case comes to us via Alvin Goldman—"Discrimination and Perceptual Knowledge," this JOURNAL, LXXIII, 20 (November 18, 1976): 771-91

[15] We should not infer from these examples that epistemic-failure worlds associated with Gettier cases are only ever *truth*-failure worlds Consider Gilbert Harman's case of the political assassination in *Thought* (Princeton: University Press, 1973), pp 143-44. In his story, Jill has good evidence for there having been an assassination, and she happens not to hear what everyone else around her has heard—a government denial of there having been any such event. If she had heard the denial, though, she might well have ceased to believe that there had been an assassination. She would at least have ceased to have the same good evidence for believing that there had been an assassination. So, in this Gettier case the threat is not to the truth component of Jill's putative knowledge. Instead, the threat is to the belief component and/or the justification component.

Nevertheless, we need not infer from this that each Gettier case's epistemic subject lacks the pertinent knowledge that *p*. Instead, section I above has provided a conceptual framework within which we may interpret those epistemic subjects as having knowledge—failable knowledge. FailK2 entails that, if the epistemic subject's belief is knowledge, it is failable knowledge. It entails this, because we have seen just now that FailK2(2) is satisfied: for each Gettier case, there are associated epistemic-failure worlds (indeed, close ones for the case's epistemic subject in relation to *p*). Hence, if FailK2(1) is also satisfied, the epistemic subject's belief is failable knowledge. And is FailK2(1) also satisfied?

The usual answer to that question would be this: "No. Within a Gettier case there is simply a lack of knowledge." To which I reply:

> But the supposed lack of knowledge is due only to the strange occurrence; and I have interpreted the strange occurrence as contributing just to FailK2(2)'s being satisfied—that is, to the failability as such, rather than to denying the knowing as such (the knowing referred to in FailK2(1)) The point is that we need not assume that the strange occurrence's presence denies the epistemic subject knowledge; instead, we may view it as affecting *how* the epistemic subject knows—such as the strength or quality or embeddedness of that knowledge, its failability rather than its presence.

> In particular, I advocate this interpretation: one's having a well-justified true belief that *p* is sufficient for one's knowing that *p*. But if there is a strange occurrence within one's context, then the knowing is failable.

In fact, we may strengthen that interpretation, by calling on FailK3 as well. Because an epistemic subject's being in a Gettier situation turns what would otherwise be some distinctive and *distant* epistemic-failure worlds into *close* ones, we may interpret such situations in the following terms:

> One's having a well-justified true belief that *p* is sufficient for one's knowing that *p*. But if there is a strange occurrence within one's context, then the knowing is very failable

This does justice to the feeling that there must be a difference in the quality of the instances of knowing in, respectively, a normal situation where there is failable knowledge that *p*, and a *Gettier* situation where there is failable knowledge that *p*. In a normal situation, there is a well-justified true belief that *p* and there is no strange occurrence. In a Gettier case, there is a well-justified true belief that *p* and there is a strange occurrence. That makes some of the epistemic-failure worlds for *x* in relation to *p* much closer than they would normally be for an epistemic subject, within *x*'s world, when knowing that *p*. Let us see why that is so (first in a general way; then by applying those general remarks to a particular example).

*The general explanation.* (i) If *x* knows that *p* within a normal (non-Gettier) situation, then none of the close—the similar—accessible worlds, as regards that knowledge of *x*'s, is an epistemic-*failure* world for *x* in relation to *p*. That is, in every close world where at least two of the three traditional components of *x*'s knowing that *p* are present, the third component is also present. Because (by our present hypothesis) *x* knows that *p* in a normal way, that normality is also present within any very close worlds where at least two of those traditional components are present, with the result—the normal result, after all—being that the third component is also present.[16] (ii) On the other hand, if *x* knows that *p* within a Gettier situation, then there are some close accessible worlds, as regards that knowledge, which are epistemic-failure worlds for *x* in relation to *p*. This is because the close accessible worlds, as regards *x*'s knowing that *p* within a Gettier case, are not simply worlds where at least two of the three traditional components of *x*'s knowing that *p* are present; they are also worlds that contain the strange occurrence from *x*'s Gettier situation. And in some of these worlds the strange occurrence's presence ensures that the third traditional component of *x*'s knowing that *p* is not present. Relative to a specific case, there need not be worlds like this for each one of the three components, but there will be some for at least one of the three. For example, the luck with which, given the strange occurrence, *x*'s good evidence for *p* and his believing that *p* has coincided in this world's Gettier situation with *p*'s being true, might be modeled by there being some close truth-failure worlds for *x* in relation to *p*—some close truth-failure worlds for *x* in relation to *p*, that is, where the strange occurrence is also present.

---

[16] Presumably, worlds where at least two of the three components are present but where *skeptical* possibilities are actualized—with the third component therefore not also being present—are rather distant ones. Worlds where a Cartesian evil genius falsifies all of one's well-justified beliefs, say, or where one's well-supported inductive predictions are falsified by a sudden Humean alteration to the world, are like that (unless, of course, our world is more like an actualization of such a world than we take it to be). Note, too, that a skeptic might regard those worlds as including ones which contain various *Gettier* situations For more on how a skeptic might claim to find skeptical potential in possible Gettier cases, see my "Gettieristic Scepticism," *Australasian Journal of Philosophy*, LXXIV, 1 (March 1996) 83-97

*A particular case.* We may now apply those general remarks to a particular case—Chisholm's Sheep Case, say. Let *S* be the proposition that there is a sheep in the field (that is, the field in front of *x* at the time in question). Is *x*'s knowing that *S* in a normal situation so much less failable than his knowing that *S* in a Gettier situation (the one described by Chisholm)?

(i) If *x* knows in a normal way that *S*, then the close worlds that determine how failable that knowledge is are these ones:

> The close worlds where *x* has a true belief that *S*.
>
> The close worlds where *x* believes that *S* and *x* has good sensory evidence *E* for *S*[17]
>
> The close worlds where *S* is true and *x* has good sensory evidence *E* for *S*.

These are the various close accessible worlds—more specifically, the close worlds where at least two of the three traditional components of *x*'s knowing that *S* are present. And in each of those worlds, too, the third traditional component of *x*'s knowing that *S* is present:

> In the close worlds where *x* has a true belief that *S*, *x* also has good sensory evidence *E* for *S*.
>
> In the close worlds where *x* believes that *S* and *x* has good sensory evidence *E* for *S*, *S* is also true.
>
> In the close worlds where *S* is true and *x* has good sensory evidence *E* for *S*, *x* also believes that *S*.

---

[17] Throughout my discussion of this case, I shall use '*E*' to denote what is qualitatively the same sensory evidence as *x* has in this world for *S*—which is to say that *E* is its appearing to *x*, on the basis of his looking at the field, that there is a sheep in that field in front of him.

These regularities obtain because those worlds are as similar to this world as is compatible with their containing at least two members of the set {*S*'s being true, *x*'s believing that *S*, *x*'s having good sensory evidence *E* for *S*}, and because in this world there is, say, a normal kind of reliability underlying the fact of *x*'s having all three members of that set. Accordingly, in those other worlds also, all three members of the set are present. The result is that none of these close worlds is an epistemic-failure world for *x* in relation to *S*.[18]

(ii) If *x* knows that *S* within a *Gettier* situation, then the close worlds that determine how failable that knowledge is are these ones:

> The close worlds where *x* has a true belief that *S*, and where the strange occurrence from *x*'s Gettier situation also obtains.
>
> The close worlds where *x* believes that *S* and *x* has good sensory evidence *E* for *S*, and where the strange occurrence from *x*'s Gettier situation also obtains.
>
> The close worlds where *S* is true and *x* has good sensory evidence *E* for *S*, and where the strange occurrence from *x*'s Gettier situation also obtains.

Let that strange occurrence be the fact that *x* is really only looking at a disguised dog (one which nevertheless seems to *x* to be a sheep).[19] Then the second of those three groups of worlds, at any rate, contains some epistemic-failure worlds for *x* in relation to *S*. Each world in that second group contains *x*'s believing that *S* and *x*'s having good sensory evidence *E* for *S*, where the presence of *E* is due, however, to *x*'s looking at what is really a *disguised dog* (one which happens to look to *x* exactly like a sheep); yet in some such worlds, *S* is not true. (In some of them, for example, the real sheep, which was out of *x*'s sight anyway, has wandered out of the field for a minute or two, including the moment when *x* is looking at the field.) Those worlds are thus close epistemic-failure worlds for *x* in relation to *p*.

But how close are they to this world—that is, to this world when it includes *x*'s knowing that *S* within a Gettier situation? In general

---

[18] Obviously, there are epistemic-failure worlds for *x* in relation to *S*. But given the normality of how—and therefore the stability with which—*x* is being supposed to know that *S* in this world, those worlds are not close ones.

[19] I am not building into my characterization of the strange occurrence the fact that there is a sheep elsewhere in the field. (On how "large" a strange occurrence needs to be, see footnote 12 above.) In any event, we could modify the criterion of closeness for the worlds we are discussing, so that they include either the strange occurrence *or* some vital part of it.

(and from FailK3), *x*'s knowledge that *S* is as failable as there are numerous and close epistemic-failure worlds for *x* in relation to *S*. And (as we have seen just now) when *x* knows that *S* within a Gettier case that knowledge is at least more failable than *x*'s normal knowledge that *S* would be: there are no close epistemic-failure worlds for *x* in relation to *S*, when *x* is in a normal situation, and there *are* some such worlds when *x* is in a Gettier situation. Surely, we may say something even stronger than that, though. Surely, we may say this:

> *x*'s knowledge that *S* within the Gettier situation is not just more failable than his normal knowledge that *S* would be; it is much more failable than his normal knowledge that *S* would be.

After all, even when *x* knows that *S* within the Gettier situation, this occurs within what is otherwise (we may assume) a normal world; and within such a world, a person would be unlikely to gain a true belief that *S* by looking at a disguised dog and seeing it, quite reasonably, as being a sheep (and by seeing no other animals in the field). Hence, while the normality of the world in general (when *x* knows that *S* in a normal situation) makes all epistemic-failure worlds for *x* in relation to *S* distant ones, the normality of the world in general apart from the strange occurrence in *x*'s situation (when *x* knows that *S* within a Gettier case) makes at least some epistemic-failure worlds for *x* in relation to *S* *close* ones. So, once more (and other things being equal), knowledge that *p* within a Gettier case is much more failable than knowledge that *p* within a normal situation.

Where does all of that leave us? I am saying—nonstandardly, I realize, but coherently, I claim—that the strange occurrence in a helpful Gettier case contributes to the epistemic subject's having knowledge, by helping to give him a well-justified true belief. By the same token, though, I am conceding that the knowledge is present only in a very fragile or finely balanced way. So, it is very failable. It is present but might easily not have been. The strange occurrence within a dangerous Gettier case, on the other hand, does not give the epistemic subject some failable knowledge. On the contrary: without the strange occurrence within a dangerous Gettier case, the epistemic subject's well-justified true belief would be knowledge in a normal way. With the strange occurrence's presence within a dangerous Gettier case, though, the knowledge is present only in a very fragile or finely balanced way. So, it is very failable. It is present but might easily not have been. Thus, what is true in this respect of helpful Gettier cases is no less true of dangerous Gettier cases.

I cannot prove that this is how Gettier cases must be interpreted. But I have done enough, I believe, to prove that the usual view of Gettier cases is not mandatory. It is standard—but not therefore true. And it is certainly not mandatory, given that we now have before us a coherent and independently motivated way to interpret each Gettier case's epistemic subject as having failable knowledge within his case.[20]

(B) Gettier cases presented us with what was an epistemological *puzzle* only on the assumption that there can be failable knowledge.[21] And the equation of knowledge with well-justified true belief had seemed like the obvious account with which to understand that assumption. But as soon as some cases—specifically, the Gettier cases—were encountered where the epistemic subjects satisfied that account while almost not doing so, the account was taken to be faulty. Each Gettier case's epistemic subject had a well-justified true belief—while almost not doing so. Hence, well-justified true belief was taken to be an inadequate account of knowing. Epistemologists were impressed by how easy it is to imagine changes to the circumstances within a Gettier case, changes which would have led to the case's epistemic subject not having the well-justified true belief he actually has.[22] They inferred that this shows that his having the well-justified true belief that $p$ which he actually has is not enough for his knowing that $p$—and hence that knowledge is *more* than a well-justified true belief.

But that did not follow at all. The fact that simple changes to the epistemic subject's situation are possible, and that in the changed situation he would not retain his well-justified true belief that *p*, does not entail that—while still within his actual, unchanged, situation—he does fail to know that *p*. It shows only that (and how) he would fail to know that *p*, if things were other than as they are in a particular respect. The fact that he would not have known that *p* does not entail that he does not know that *p*. Actual knowing does not entail counterfactual knowing.

---

[20] Dare I mention, too, how hard—or worse—it has been for epistemologists to agree on an analysis of knowledge while also accepting, as they do, that in Gettier cases there is a lack of knowledge? My speculation is that the various analyses are mistaken to at least the extent that the acceptance underlying them—the acceptance that in Gettier cases there is a lack of knowledge—is mistaken. At any rate, other theoretical options should be formulated and discussed; I am offering one such option

[21] If knowledge is required to be *in*failable, there is no independent puzzle about why there is no knowledge within Gettier cases The knowledge would be absent because the epistemic subject's justification, say, is less than perfect; and that would be that. Gettier cases are puzzling only if we believe that knowledge need not be infailable—for instance, it can be fallible (For then our best initial theory of such knowledge—which allows it to be fallibly-but-well-justified true belief, say—is satisfied, and accordingly we should think that knowledge is present. Yet epistemologists have not thought that it is present. Thus, they have generated a puzzle for themselves.)

[22] In helpful Gettier cases, the change is the strange occurrence's being absent (for example, the sheep's not being in the field). In dangerous Gettier cases, the change is the strange occurrence's "carrying out" its threat (for example, Henry's actually being misled by a fake barn).

Admittedly, only a small alteration to the context is needed to bring about the imagined counterfactual situation—and unarguably the epistemic subject does lack the knowledge within that altered situation. Thus, the epistemic subject could easily fail to have that knowledge. And this might be thought to show that the alternative situation is similar enough to the actual situation to be epistemically relevant to the actual situation. It might be thought that therefore one's knowing within the actual situation does depend on one's knowing within the counterfactual situation—so that, because one fails to know in the latter situation, one also fails to know in the former situation. It might be thought that the epistemic subject does fail to know, because he could easily fail to know. It might be thought that this is why there is a lack of knowledge in the actual situation—the Gettier situation.

Well, that is the usual way of interpreting Gettier cases. But we have found another way to do so—and now I shall extend this alternative interpretation a little. If we are to understand the nature of failable knowing, we could say instead (and this is my key observation here) that the closeness of the alternative possible situation—in which the epistemic subject definitely does fail to know that *p*—entails only that the epistemic subject's actual situation is almost one where he has that lack of knowledge. Because he could so easily fail to know, he does almost fail to know. His knowing is very failable, because he almost fails to know. The closeness of the alternative possible situation in which he does fail to know does not entail that the actual situation is one where there is a lack of knowledge. After all, in fact the small alteration to his actual situation has not occurred. And even if we feel that the person is just lucky that it did not occur, the fact remains that it did not occur. This is a fact which should be reflected in our assessment of the epistemic situation. It certainly plays a part in mine:

> Within each Gettier case, the epistemic subject knows, because he has a well-justified true belief. But he is very lucky to have all of that (and thereby the knowledge). This luck is a mark of how very failably he knows

> within the Gettier case. He does know—even though he almost failed to do so. He does know—even though he is *very* close to not doing so.

Consider an analogous situation. Maybe there is luck involved in life's ever having come to exist in this universe. There are so many possible—and slight—differences to the universe's initial conditions which, had they been actual, would not have led to the creation of life. So, by having the initial conditions it had, the universe was close to having different initial conditions—and thereby to not containing life. Nevertheless, it did not have those different initial conditions. Most of us do not say that consequently there are no natural laws within the universe. There are, in spite of the luck involved in their obtaining. And failable knowing within a Gettier case is like that. In each such case, the epistemic subject is very close to failing to know that *p*. He almost fails to know that *p*. And his being so close to failing to know that *p* is what misleads people into thinking that he does fail to know that *p*. Instead, we may say that his being so close to not knowing that *p* is simply part of his knowing very failably that *p*.

(C) That interpretation of Gettier cases clashes with standard "intuitions" about the cases. But intuitions without developed theories to accompany them are blind; and as yet there is no deeply happy marriage between those intuitions and a theory of failable knowledge. (More limitedly, there is not even a union like that between those intuitions and theory of fallible knowledge.) What is more, we should wonder what underdeveloped theoretical commitments are lurking behind those "intuitions." Epistemological reactions such as those "intuitions" are not givens. Might they be mistaken? Of course they might. Might they be based upon some false or misleading theoretical views that are unreflectively being taken for granted by epistemologists? Of course they might.

For example, epistemologists should ask themselves whether their so-called "intuitions" about Gettier cases might have relied implicitly upon the following thesis:

> KNAF: a person knows that *p*, only if it is not the case that he almost fails to know that *p* That is, knowing must not be a "near thing." It must not be lucky. It must not almost not occur (if it is to occur). This is so, even for failable knowledge.

Undoubtedly, an adherence to KNAF would lead to the usual view of Gettier cases (because within a Gettier case the epistemic subject does at least come close to failing to know—and because, according to KNAF, that suffices for failing to know, which *is* the usual view of what happens within a Gettier case). And it is not hard to imagine how an epistemologist could be attracted to that thesis, KNAF. It would flow smoothly, if there was also an attraction to this thesis:

> What is not hard to achieve is not easy not to achieve. And knowledge is not hard to achieve. It is absent if, in the circumstances, it could easily have been absent.

Presumably, the idea behind such thinking is that there is something particularly "everyday"—and correlatively robust or resilient—about knowing. Such thinking might well seem plausible to many epistemologists.

Nevertheless, it is fundamentally mistaken thinking. Failable knowing can be achieved either robustly or not, in an "everyday" way or not; either way, the result remains failable knowing. Indeed, presumably its being achieved less than robustly—in an unusual way—is more likely, given its failability; indeed again, that can be part of its failability. Knowing can be at least somewhat lucky knowing—perhaps more so when it is failable knowing. It can be nonstandard or unusual knowing—again, perhaps more so when it is failable knowing. The usual reaction to Gettier cases, it seems to me, confuses failable knowing with failable and robust knowing. It confuses failable knowing with failable and wholly standard or usual knowing. The knowing inside a Gettier case is failable—but not failable and robust, or failable and wholly standard or usual. The fact of some knowing's being at least somewhat lucky, though, does not entail that it is not knowledge.[23]

In that way, it is open to us to view strange occurrences within Gettier cases as mainly affecting the quality of the knowing. First and foremost, they are a mark of its failability—even a mark of its marked failability. Within a Gettier case, the epistemic subject knows—but almost did not. Within a Gettier case, the knowledge is more failable than most of our failable knowledge. (That is because Gettier cases, unlike most situations in which we have failable knowledge, contain strange occurrences.) Epistemologists have misinterpreted Gettier cases, in part by assuming that there is only one coherent way *to* interpret the cases. There has been a general desire to interpret knowing less demandingly (so that people can have knowledge comparatively easily)—while nevertheless also adopting what is really a more demanding conception of knowing (so that people in Gettier cases lack knowledge). But that is simply an incoherent set of attitudes. One of them has to be discarded. I am outlining a way to do that, by interpreting knowing less demandingly (and, correlatively, allowing that people in Gettier cases do not lack knowledge).

---

[23] Any temptation to deny that this is so should be tempered by a realization of how demanding a conception of knowledge is implied by such a denial. To insist that there can never be any luck in someone's knowing that *p* is to be an infailabilist; and an infailabilist's conception of knowledge is even more demanding than an infallibilist's. From FalK, an infallibilist says that knowledge's justification component entails the truth of its belief component—which is to say that, if the belief component and the justification component are present, then the truth component has to be present, too. But an infailabilist says that, if any two of the three components are present, then the third one has to be present. So, an infallibilist conception places upon knowledge only one of the three requirements placed upon knowledge by an infailabilist conception. An infallibilist conception is thus much less demanding than an infailabilist one.

Is it too undemanding an interpretation, though? Does it imply, for instance, that we no longer need to have good justification at all for *p* if we are to know failably that *p*? In particular, does it imply that this is true of the very failable knowledge which, I have argued, is present within Gettier cases? Here is an objection to my interpretation, based upon the suspicion that it does have that implication:

> The interpretation in this paper claims that there is very failable knowledge that *p* within Gettier cases—that there is knowledge present but only somewhat luckily. But the luck that is present within Gettier cases is a feature, first and foremost, of how the true belief that *p* is present. And this suggests that knowledge need not include justification: if knowledge can be as failable as the interpretation claims it to be within Gettier cases, then knowledge can be a mere true belief—with the luck in the true belief's presence constituting the failability of the knowledge. The interpretation is thus congenial to relinquishing the traditional justification condition on knowing.

But that objection to my interpretation fails. Within many a Gettier case there is indeed some luck; and we might well wish to describe it as there being some luck in the true belief's having been gained. Such a description would be misleadingly brief, though. A fuller, and more accurate, description would say that the luck is a matter of the belief's being true, *given* how the belief has been reached—where in fact it has been reached by the epistemic subject's relying on his evidence. For example, let us suppose that we are inclined to view Chisholm's Sheep Case as one whose epistemic subject *x* is rather lucky to have a true belief that *S* (that is, the true belief that there is a sheep in the field). Of what, precisely, is that luck a feature? Is the luck simply a matter of *S*'s being luckily true, say? I do not think so: there is a sheep in the field, and perfectly normally so, with *S*'s truth as such thereby being no less normal. For the same reason, too, the luck is not strictly a matter of *x*'s *belief* that *S* being luckily true. Instead, the luck pertains to *x*'s having the true belief that *S* on the basis of his good sensory evidence *E*. There are different ways in which we might try to explicate that accidentality. One popular way is to say that because $E$ is misleading $x$ into believing that there is a sheep in front of him (remember that $E$ is being produced within $x$ by a dog who is disguised to look like a sheep), $x$ is somewhat lucky then to gain the true belief that $S$ on the basis of $E$. Yet, however we explicate it, the fact remains that the luck within the case is a relationship—one about whose more precise nature there is epistemological debate—between (i) $S$'s being true and/or $x$'s believing that $S$, *and* (ii) $x$'s relying on $E$ in gaining that belief. So, even when a Gettier case's epistemic subject has his true belief somewhat luckily, and even if (as I have argued) that true belief is correlatively failable knowledge, we need not infer from this that knowledge does not have to include a justification component. This is because the luck in the true belief's presence within the Gettier case is only present insofar as the good evidence is also present and insofar as the true belief is based on that evidence. In other words, the luck is not really a feature of the *true belief's* presence as such; it is a feature of the justified true belief's presence. And thus the objection fails. Of course, its failure does not entail, in turn, that all knowledge does need to include justification. It does entail, though, that this paper's interpretation of Gettier cases in particular does not entail that knowledge need not include justification.[34]

(D) Let us suppose that a chosen few of the brush strokes constituting the *Mona Lisa* were so integral to it that if they had been different, the painting would have suffered greatly. Call these *the strokes*. If they had been absent, the painting would have lacked the sensitivity and accuracy which (I assume) it has. Naturally, even now it is not a perfect painting. It is fine; yet it would not have been fine if the strokes had been absent. Moreover, they might easily have been absent, with Leonardo da Vinci having chosen between them and some "competing" strokes, say. In fact, however, they were not absent. As a result, the picture is the fine portrait that it is.

---

[34] In saying this, I am not denying that there could be some knowledge that does not include justification; I am saying only that the knowledge within Gettier cases is not like that, in spite of that knowledge being as failable as it is. Recently, Crispin Sartwell has argued for the even stronger thesis that knowledge is mere true belief (and hence, strictly speaking, never includes justification)—"Knowledge Is Merely True Belief," *American Philosophical Quarterly*, XXVIII, 2 (April 1991): 157-65, and "Why Knowledge Is Merely True Belief," this JOURNAL, LXXXIX, 4 (April 1992): 167-80. Elsewhere (in part of a manuscript, "Good Knowledge, Bad Knowledge"), I have rejected Sartwell's position, while arguing that there *could* be knowledge which is unjustified. That argument is independent of the one here But if—on independent grounds—we conclude that there could be instances of knowledge that do not include justification, then we may extend my analysis of failability correlatively: the pertinent epistemic-failure worlds for those cases of unjustified knowledge would be ones where one and only one of the truth component and the belief component is present.

But imagine that initially the strokes had been absent, with "competitors" taking their place—and that da Vinci, bleary-eyed and tired, and barely aware of what he was doing, proceeded to replace the competitors with the strokes. The *Mona Lisa* as we know it was the product of his doing so. But his doing so also means the painting's having been produced in a way that is analogous to the production of a well-justified true belief within a Gettier case. (The strokes having been absent, and then all-but-accidentally being included after all, is this case's analogue of a strange occurrence within a Gettier case—specifically, a helpful Gettier case.) The painting is accurate (which is like a belief's being true); its accuracy is sensitively rendered (which is like a belief's being justified); and a nontrivial kind of luck (which is like a strange occurrence's being present) contributed to the painting's ending up with that combination of accuracy and sensitivity. Because of the luck, the painting almost failed to be as fine as it is; even so, that is what it is—fine, very fine.

An analogous moral pertains to knowing failably. A belief's being produced, in part, by a strange occurrence such as is present in helpful Gettier cases does not prevent its being failable knowledge (so long as the belief is true and well justified). We may interpret the strange occurrence, instead, as part of why the knowledge—the well-justified true belief—which it helps to produce is only failable. The belief's nevertheless being knowledge is like the *Mona Lisa*'s being a fine and accurate painting even if its combination of accuracy and sensitivity was to have been partly due to luck. The failability of the painting's high quality would be partly due to the luck: it is a fine painting, but (in the imagined situation) it almost was not.

(E) Sometimes, boldness is needed when claiming or attributing knowledge that is failable. My way of speaking about knowing makes clear why that is so:

> A person knows failably that *p*, if and only if he is at least a little close to not knowing that *p*. He knows very failably that *p*, if and only if he is very close to not knowing that *p*. (Equivalently: a person knows very failably that *p*, if and only if he almost fails to know that *p*.)

To attribute failable knowledge is to attribute knowledge that might have been absent. The more failable it is, the less confident we might be that it is knowledge. But this should not worry us; in particular, it should not lead to our precipitately assuming that knowledge that *p* is absent whenever our confidence in its being present begins to ebb away (such as occurs with Gettier cases). We need to keep in mind that—other things being equal—the more failable a piece of knowing is, the less confident we should be that it is knowledge.

That might sound mistaken: "Surely, the more failable a piece of knowledge is, the more confident we should be that it is knowledge. After all, the more failable it is, the lower a justificatory standard needs to be satisfied by it, for example." Not so (I say): the less failable a case of knowing is, the further away it is from not being knowledge; accordingly, the more confident we should be that it is knowledge. If one knows *infailably* that *p*, one is maximally far away from not knowing that *p*. If ever one should be confident that one knows that *p*, it is in this case.[25]

Naturally or conventionally (I shall not hazard a view as to which), knowing has boundaries. Definitionally, the lower boundary of failable knowing divides (i) not knowing while almost knowing, from (ii) knowing while almost not knowing. Why should we not also distinguish—within the category of failable knowing—between (ii) knowing while almost not knowing, and (iii) knowing while not even almost not knowing? Part of my thesis is that we should be on the alert for instances of that distinction. And the distinction between (a) the epistemic subject's well-justified true belief within a Gettier case, and (b) most normal failable instances of knowing, is such an instance: (a) stands to (b) as (ii) stands to (iii). Moreover, why should we not describe cases of (ii) as being cases of more failable knowing than are cases of (iii)—cases of knowing which are closer to not being cases of knowing? Part of my thesis is that we should do so, since it marks a useful way of reflecting more carefully on various people's epistemic situations. The knowledge within a Gettier case, say, is more failable than most. When one is in a Gettier case, one is very close to not knowing. That being so, it is not surprising that beliefs within Gettier cases have been mistaken for instances of not knowing. When working at the boundary of a concept, mistakes occur easily; this is why philosophical counterexamples can be puzzling. Gettier cases are a perfect example of this—of philosophical counterexamples that have generated puzzlement in that way. In particular, they are cases that test the lower boundary of the concept of failable knowing.

---

[25] What one is entitled not to be so confident about in that event is the infailability as such of one's infailable knowledge that *p*. To be confident that some knowledge that *p* is infailable, as against all-but-infailable, is not as warranted as is the fact that, *either* way, one knows that *p*—in short, that one knows simpliciter that *p* (We shall disagree with this only if we believe that knowledge must be infailable, hence if we believe that all knowledge simpliciter is infailable knowledge. But that is not a belief I am modeling here.)

The problem has been that epistemologists—through the mistaken kind of caution discussed in subsection (C), I surmise—have placed the cases on the wrong side of that boundary. Still, this was almost not a mistake by epistemologists. We can say that much for it, I suppose. A careful initial reaction to Gettier cases would be to say that their epistemic subjects either lack knowledge or have it but only just. Some knowing has to be at the border, between knowing and not knowing—that is, some knowing is almost not knowing. It would be easy to mistake such knowing for not knowing. And this is precisely what, it seems to me, epistemologists have done in their responses to Gettier cases. At the very least, therefore, epistemologists should abandon their customary confident belief that in Gettier cases knowledge is absent—and that this is the only reasonable interpretation of the cases. It is not.

STEPHEN HETHERINGTON

University of New South Wales

## BARGAINING WITH NEIGHBORS: IS JUSTICE CONTAGIOUS?

What is justice? The question is harder to answer in some cases than in others. We focus on the easiest case of distributive justice. Two individuals are to decide how to distribute a windfall of a certain amount of money. Neither is especially entitled, or especially needy, or especially anything—their positions are entirely symmetric. Their utilities derived from the distribution may be taken, for all intents and purposes, simply as the amount of money received. If they cannot decide, the money remains undistributed and neither gets any. The essence of the situation is captured in the simplest version of a bargaining game devised by John Nash.[1] Each person decides on a bottom-line demand. If those demands do not jointly exceed the windfall, then each person gets his demand; if not, no one gets anything. This game is often simply called *divide-the-dollar*.

In the ideal simple case, the question of distributive justice can be decided by two principles:

> *Optimality*: a distribution is not just if, under an alternative distribution, all recipients would be better off.
>
> *Equity*: if the position of the recipients is symmetric, then the distribution should be symmetric. That is to say, it does not vary when we switch the recipients.

Since we stipulate that the position of the two individuals is symmetric, equity requires that the just distribution must give them the same amount of money. Optimality then rules out such unlikely schemes as giving each one dime and throwing the rest away—each must get half the money.

There is nothing new about our two principles. Equity is the simplest consequence of the theory of distributive justice in Aristotle's *Politics*. It is a consequence of Immanuel Kant's categorical imperative. Utilitarians tend to stress optimality, but are not completely insensitive to equity. Optimality and equity are the two most uncontroversial requirements in Nash's axiomatic treatment of bargaining. If you ask people to judge the just distribution, their answers show that optimality and equity are powerful operative principles.[2] So, although nothing much hangs on it, we shall feel

---

[1] "The Bargaining Problem," *Econometrica*, XVIII (1950): 155-62.

[2] Menachem Yaari and Maya Bar-Hillel, "On Dividing Justly," *Social Choice and Welfare*, I (1981): 1-24.

0022-362X/99/9611/588–98

free to use moral language and to call the equal split *fair division* in divide-the-dollar.

## I. RATIONALITY, BEHAVIOR, EVOLUTION

Two rational agents play the divide-the-dollar game. Their rationality is common knowledge. What do they do? The answer that game theory gives us is that *any* combination of demands is compatible with these assumptions. For example, Jack may demand ninety percent thinking that Jill will only demand ten percent on the assumption that Jill thinks that Jack will demand ninety percent and so forth, while Jill demands seventy-five percent thinking that Jack will demand twenty-five percent on the assumption that Jack thinks that Jill will demand seventy-five percent and so forth. *Any* pair of demands is *rationalizable*, in that it can be supported by a hierarchy of conjectures for each player, compatible with common knowledge of rationality. In the example given, these conjectures are quite mistaken.

Suppose we add the assumption that each agent somehow knows what the other will demand. Then any combination of demands that total the whole sum to be divided is still possible. For example, suppose that Jack demands ninety percent knowing that Jill will demand ten percent and Jill demands ten percent knowing that Jack will demand ninety percent. Then each player is maximizing payoff given the demand of the other. That is to say that this is a Nash equilibrium of divide-the-dollar. If the dollar is infinitely divisible, then there are an infinite number of such equilibria.

If experimental game theorists have people actually play divide-the-dollar, they *always* split equally.[1] This is not always true in more complicated bargaining experiments where there are salient asymmetries, but it is true in divide-the-dollar. Rational-choice theory has no explanation of this phenomenon. It appears that the experimental subjects are using norms of justice to select a particular Nash equilibrium of the game. But what account can we give for the existence of these norms?

Evolutionary game theory (reading 'evolution' as cultural evolution) promises an explanation, but the promise is only partially fulfilled. Demand-half is the only evolutionarily stable strategy in divide-the-dollar.[4] It is the only strategy such that, if the whole population played that strategy, no small group of innovators, or "mutants," playing a different strategy could achieve an average payoff at least as great as the natives. If we could be sure that this unique evolutionarily stable strategy would always take over the population, the problem would be solved.

---

[1] Rudy V Nydegger and Guillermo Owen, "Two-Person Bargaining: An Experimental Test of the Nash Axioms," *International Journal of Game Theory*, III (1974): 239-50; Alvin Roth and Michael Malouf, "Game Theoretic Models and the Role of Information in Bargaining," *Psychological Review*, LXXXVI (1979): 574-94; John Van Huyck, Raymond Battalio, Sondip Mathur, Patsy Van Huyck, and Andreas Ortmann, "On the Origin of Convention: Evidence From Symmetric Bargaining Games," *International Journal of Game Theory*, XXIV (1995) 187-212.

But we cannot be sure that this will happen. There are states of the population which are evolutionarily stable where some fraction of the population makes one demand and some fraction makes another. The state where half the population demands one third and half the population demands two thirds is such an evolutionarily stable polymorphism of the population. So is the state where two thirds of the population demands forty percent and one third of the population demands sixty percent We can think of these as pitfalls along the evolutionary road to justice.

How important are these polymorphisms? To what extent do they compromise the evolutionary explanation of the egalitarian norm? We cannot begin to answer these questions without explicitly modeling the evolutionary dynamics and investigating the size of their basins of attraction.

## II. BARGAINING WITH STRANGERS

The most widely studied dynamic evolutionary model is a model of interactions with strangers. Suppose that individuals are paired at random from a very large population to play the bargaining game. We assume that the probability of meeting a strategy can be taken as the proportion of the population that has that strategy. The population proportions evolve according to the replicator dynamics. The proportion of the population using a strategy in the next generation is the proportion playing that strategy in the current generation mutiplied by a *fitness factor*. This fitness factor is just the ratio of the average payoff to this strategy to the average payoff in the whole population.[5] Strategies that do better than average grow; those which do worse than average shrink. This dynamic arose in biology as a model of asexual reproduction, but more to the point here, it also has a cultural evolutionary interpretation where strategies are imitated in proportion to their success.[8]

---

[4] Robert Sugden, *The Economics of Rights, Cooperation, and Welfare* (New York: Blackwell, 1986).

[5] This is the discrete time version of the replicator dynamics, which is most relevant in comparison to the alternative *bargaining-with-neighbors* dynamics considered here. There is also a continuous time version. As comprehensive references, see Josef Hofbauer and Karl Sigmund, *The Theory of Evolution and Dynamical Systems* (New York: Cambridge, 1988); Jorgen W. Weibull, *Evolutionary Game Theory* (Cambridge: MIT, 1995); Larry Samuelson, *Evolutionary Games and Equilibrium Selection* (Cambridge: MIT, 1997).

The basins of attraction of these polymorphic pitfalls are not negligible. A realistic version of divide-the-dollar will have some finite number of strategies instead of the infinite number that we get from the idealization of infinite divisibility. For a finite number of strategies, the size of a basin of attraction of a population state makes straightforward sense. It can be estimated by computer simulations. We can consider coarse-grained or fine-grained versions of divide-the-dollar; we can divide a stack of quarters, or of dimes, or of pennies. Some results of simulations persist across a range of different granularities. Equal division always has the largest basin of attraction and it is always greater than the basins of attractions of all the polymorphic pitfalls combined. If you choose an initial population state at random, it is more probable than not that the replicator dynamics will converge to a state of fixation of demand-half. Simulation results range between fifty-seven and sixty-three percent of the initial points going to fair division. The next largest basin of attraction is always that closest to the equal split: for example, the four-six polymorphism in the case of dividing a stack of ten dimes and the forty-nine/fifty-one polymorphism in the case of dividing a stack of one-hundred pennies. The rest of the polymorphic equilibria follow the general rule—the closer to fair division, the larger the basin of attraction.

For example, the results running the discrete replicator dynamics to convergence and repeating the process 100,000 times on the game of dividing ten dimes are given in table 1.

| | |
|---|---|
| Fair Division | 62,209 |
| 4-6 Polymorphism | 27,469 |
| 3-7 Polymorphism | 8,801 |
| 2-7 polymorphism | 1,483 |
| 1-9 Polymorphism | 38 |
| 0-10 Polymorphism | 0 |

Table 1: Convergence results for replicator dynamics - 100,000 trials

---

[8] Jonas Björnerstedt and Jörgen Weibull, "Nash Equilibrium and Evolution by Imitation," in Kenneth J. Arrow et alia, eds., *The Rational Foundations of Economic Behavior* (New York: Macmillan, 1996), pp. 155-71, Karl Schlag, "Why Imitate, and If So How?" Discussion Paper B-361 (University of Bonn, Germany, 1996).

The projected evolutionary explanation seems to fall somewhat short. The best we might say on the basis of pure replicator dynamics is that fixation of fair division is more likely than not, and that polymorphisms far from fair division are quite unlikely.

We can say something more if we inject a little bit of probability into the model. Suppose that every once and a while a member of the population just picks a strategy at random and tries it out—perhaps as an experiment, perhaps just as a mistake. Suppose we are at a polymorphic equilibrium—for instance, the four-six equilibrium in the problem of dividing ten dimes. If there is some fixed probability of an experiment (or mistake), and if experiments are independent, and if we wait long enough, there will be enough experiments of the right kind to kick the population out of the basin of attraction of the four-six polymorphism and into the basin of attraction of fair division and the evolutionary dynamics will carry fair division to fixation. Eventually, experiments or mistakes will kick the population out of the basin of attraction of fair division, but we should expect to wait much longer for this to happen. In the long run, the system will spend most of its time in the fair-division equilibrium. Peyton Young[7] showed that, if we take the limit as the probability of someone experimenting gets smaller and smaller, the ratio of time spent in fair division approaches one. In his terminology, fair division is the *stochastically stable equilibrium* of this bargaining game.

This explanation gets us a probability arbitrarily close to one of finding a fair-division equilibrium if we are willing to wait an arbitrarily long time. But one may well be dissatisfied with an explanation that lives at infinity. (Putting the limiting analysis to one side, pick some plausible probability of experimentation or mistake and ask yourself how long you would expect it to take in a population of 10,000, for 1,334 demand-six types simultaneously to try out being demand-five types and thus kick the population out of the basin of attraction of the four-six polymorphism and into the basin of attraction of fair division.[8]) The evolutionary explanation still seems less than compelling.

---

[7] "An Evolutionary Model of Bargaining," *Journal of Economic Theory*, LIX (1993): 145-68, and "The Evolution of Conventions," *Econometrica*, LXI (1993): 57-94; and Dean Foster and Young, "Stochastic Evolutionary Game Dynamics," *Theoretical Population Biology*, XXXVIII (1990): 219-32.

[8] For discussion of expected waiting times, see Glenn Ellison, "Learning, Local Interaction and Coordination," *Econometrica*, LXI (1993): 1047-71; and Robert Axtell, Joshua M. Epstein, and H. Peyton Young, "The Emergence of Economic Classes in an Agent-Based Bargaining Model," preprint (Brookings Institution, 1999).

## III BARGAINING WITH NEIGHBORS

The model of random encounters in an infinite population that motivates the replicator dynamics may not be the right model. Suppose interactions are with neighbors. Some investigations of cellular automaton models of prisoner's dilemma and a few other games show that interactions with neighbors may produce dynamical behavior quite different from that generated by interactions with strangers.[9] Bargaining games with neighbors have not, to the best of our knowledge, previously been studied.

Here, we investigate a population of 10,000 arranged on a one hundred by one hundred square lattice. As the neighbors of an individual in the interior of the lattice, we take the eight individuals to the N, NE, E, SE, S, SW, W, NW. This is called the Moore(8) neighborhood in the cellular automaton literature.[10] The dynamics is driven by imitation. Individuals imitate the most successful person in the neighborhood. A generation—an iteration of the discrete dynamics—has two stages. First, each individual plays the divide-ten-dimes game with each of her neighbors using her current strategy. Summing the payoffs gives her current success level. Then each player looks around her neighborhood and changes her current strategy by imitating her most successful neighbor, providing that her most successful neighbor is more successful than she is; otherwise, she does not switch strategies. (Ties are broken by a coin flip.)

In initial trials of this model, fair division *always* went to fixation. This cannot be a universal law, since you can design "rigged" configurations where a few demand-one-half players are, for example, placed in a population of demand-four and demand-six players with the latter so arranged that there is a demand-six type who is the most successful player in the neighborhood of every demand-one-half player. Start enough simulations at random starting points and sooner or later you will start at one of these.

We ran a large simulation starting repeatedly at randomly chosen starting points. Fair division went to fixation in more than ninety-nine point five percent of the trials. The cases where it did not were all cases where the initial population of 10,000 contained fewer than seventeen demand-one-half players. Furthermore, convergence was remarkably quick. Mean time to fixation of fair division was about sixteen generations. This may be compared with a mean time to convergence[11] in discrete replicator dynamics of forty-six generations, and with the ultra-long-run character of stochastically stable equilibrium.

---

[9] Gregory B. Pollack, "Evolutionary Stability on a Viscous Lattice," *Social Networks*, XI (1989): 175-212; Martin A. Nowak and Robert M. May, "Evolutionary Games and Spatial Chaos," *Nature*, CCCLIX (1992): 826-29, Kristian Lindgren and Mats Nordahl, "Evolutionary Dynamics in Spatial Games," *Physica D*, LXXV (1994): 292-309; Luca Anderlini and Antonella Ianni, "Learning on a Torus," in Cristina Bicchieri, Richard Jeffrey, and Brian Skryms, eds., *The Dynamics of Norms* (New York: Cambridge, 1997), pp. 87-107.

[10] We find that behavior is not much different if we use the von Neumann neighborhood: N, S, E, W, or a larger Moore neighborhood.

It is possible to exclude fair division from the possible initial strategies in the divide-ten-dimes game and start at random starting points that include the rest. If we do this, all strategies other than demand-four dimes and demand-six dimes are eliminated and the four-six polymorphic population falls into a "blinking" cycle of period two. If we then turn on a little bit of random experimentation or "mutation" allowing the possibility of demand-five, we find that as soon as a very small clump of demand-five players arises, it systematically grows until it takes over the whole population—as illustrated in figure 1. *Justice is contagious.*[12]

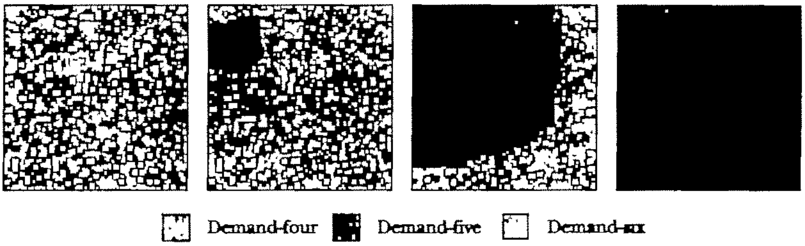


Figure 1: The steady advance of fair division

## IV ROBUSTNESS

The bargaining-with-neighbors model of the last section differs from the bargaining with strangers model in more than one way. Might the difference in behavior that we have just described be due to the imitate-the-most-successful dynamics rather than the neighbor effect? To answer this question, we ran simulations varying these factors independently.

We consider both fixed and random neighborhoods. The models using fixed neighborhoods use the Moore (8) neighborhood described above. In the alternative random-neighborhood model, each generation a new set of "neighbors" is chosen at random from the population for each individual. That is to say, these are neighborhoods of *strangers*.

---

[11] At .9999 level to keep things comparable.

[12] Ellison (*op. cit.*) found such contagion effects in local interaction of players arranged on a circle and playing pure coordination games

We investigated two alternative dynamics. One imitates the most successful neighbor as in our bargaining-with-neighbors model. The other tempers the all-or-nothing character of imitate-the-best. Under it, an individual imitates one of the strategies in its neighborhood that is more successful than it (if there are any) with relative probability proportional to their success in the neighborhood. This is a move in the direction of the replicator dynamics.

| | Bargaining with Neighbors | | Bargaining with Strangers | |
|---|---|---|---|---|
| | *A* | *B* | *C* | *D* |
| 0-10 | 0 | 0 | 0 | 0 |
| 1-9 | 0 | 0 | 0 | 0 |
| 2-8 | 0 | 0 | 54 | 57 |
| 3-7 | 0 | 0 | 550 | 556 |
| 4-6 | 26 | 26 | 2560 | 2418 |
| fair | 9972 | 9973 | 6833 | 6964 |

Table 2: Convergence results for five series of 10,000 trials

In table 2, *A* and *B* are bargaining with neighbors, with imitate-the-best-neighbor and imitate-with-probability-proportional-to-success dynamics, respectively. The results are barely distinguishable. *C* and *D* are the random-neighborhood models corresponding to *A* and *B*, respectively. These results are much closer to those given for the replicator dynamics in table 1. The dramatic difference in convergence to fair division between our two models is due to the structure of interaction with neighbors.

## V ANALYSIS

Why is justice contagious? A strategy is contagious if an initial "patch" of that strategy will extend to larger and larger patches. The key to contagion of a strategy is interaction along the edges of the patch, since in the interior the strategy can only imitate itself.[13]

---

[13] For this reason, "frontier advantage" is used to define an unbeatable strategy in Illan Eshel, Emilia Sansone, and Avner Shaked, "Evolutionary Dynamics of Populations with a Local Interaction Structure," working paper (University of Bonn, 1996).

Consider an edge with demand-five players on one side, and players playing the complementary strategies of one of the polymorphisms on the other. Since the second rank of demand-five players always meet their own kind, they each get a total payoff of forty from their eight neighbors. Players in the first rank will therefore imitate them unless a neighbor from the polymorphism gets a higher payoff. The low strategy in a polymorphic pair cannot get a higher payoff. So if demand-five is to be replaced at all, it must be by the high strategy of one of the polymorphic pairs.

In the four-six polymorphism—the polymorphism with the greatest basin of attraction in the replicator dynamics—this simply cannot happen, even in the most favorable circumstances. Suppose that we have someone playing demand-six in the first rank of the polymorphism, surrounded on his own side by compatible demand-four players to boost his payoff to the maximum possible.[14] Since he is in the first rank, he faces three incompatible demand-five neighbors. He has a total payoff of thirty while his demand-five neighbors have a total payoff of thirty-five. Demand-five begins an inexorable march forward as illustrated in figure 2. (The pattern is assumed to extend in all directions for the computation of payoffs of players at the periphery of what is shown in the figure.)

| Initial | | Iteration 1 |
|---|---|---|
| 5544 | | 5554 |
| 5544 | | 5554 |
| 5564 | => | 5554 |
| 5544 | | 5554 |
| 5544 | | 5554 |

Figure 2: fair division versus four-six polymorphism

---

[14] In situating the high strategy of the polymorphic pair in a sea of low-strategy players, we are creating the best-case scenario for the advancement of the polymorphism into the patch of demand-five players.

If we choose a polymorphism that is more extreme, however, it is possible for the high strategy to replace some demand-five players for a while. Consider the one-nine polymorphism, with a front line demand-nine player backed by compatible demand-one neighbors. The demand-nine player gets a total payoff of forty-five—more than anyone else—and thus is imitated by all his neighbors. This is shown in the first transition in figure 3.

| Initial | | Iteration 1 | | Iteration 2 | | Iteration 3 |
|---|---|---|---|---|---|---|
| 55111 | | 55511 | | 55551 | | 55555 |
| 55111 | | 55511 | | 55559 | | 55559 |
| 55111 | => | 59991 | => | 55999 | => | 55599 |
| 55911 | | 59991 | | 55999 | | 55599 |
| 55111 | | 59991 | | 55999 | | 55599 |
| 55111 | | 55511 | | 55559 | | 55559 |
| 55111 | | 55511 | | 55551 | | 55555 |

Figure 3: fair division versus one-nine polymorphism

But the success of the demand-nine strategy is its own undoing. In a cluster of demand-nine strategies, it meets itself too often and does not do so well. In the second transition, demand-five has more than regained its lost territory, and in the third transition it has solidly advanced into one-nine territory.

Analysis of the interaction along an edge between demand-five and other polymorphisms is similar to one of the cases analyzed here.[18] Either the polymorphism cannot advance at all, or the advance creates the conditions for its immediate reversal. A complete analysis of this complex system is something that we cannot offer. But the foregoing does offer some analytic insight into the contagious dynamics of equal division in "bargaining with neighbors."

## VI. CONCLUSION

Sometimes we bargain with neighbors, sometimes with strangers. The dynamics of the two sorts of interaction are quite different. In the bargaining game considered here, bargaining with strangers—modeled by the replicator dynamics—leads to fair division from a randomly chosen starting point about sixty percent of the time. Fair division becomes the unique answer in bargaining with strangers if we change the question to that of stochastic stability in the ultra-long-run. But long expected waiting times call the explanatory significance of the stochastic stability result into question.

---

[18] With some minor complications involving ties.

Bargaining with neighbors almost always converges to fair division and convergence is remarkably rapid. In bargaining with neighbors, the local interaction generates clusters of those strategies which are locally successful. Clustering and local interaction together produce positive correlation between like strategies. As noted elsewhere,[16] positive correlation favors fair division over the polymorphisms. In bargaining with neighbors, this positive correlation is not something externally imposed but rather an unavoidable consequence of the dynamics of local interaction. As a consequence, once a small group demand-half players is formed, justice becomes contagious and rapidly takes over the entire population.

Both bargaining with strangers and bargaining with neighbors are artificial abstractions. In initial phases of human cultural evolution, bargaining with neighbors may be a closer approximation to the actual situation than bargaining with strangers. The dynamics of bargaining with neighbors strengthens the evolutionary explanation of the norm of fair division.

JASON ALEXANDER
BRIAN SKYRMS

University of California/Irvine

---

[16] Skyrms, "Sex and Justice," this JOURNAL, XCI, 6 (June 1994): 305-20 and *Evolution of the Social Contract* (New York: Cambridge, 1996).

THE JOURNAL OF PHILOSOPHY
VOLUME XCVI, NO. 12, DECEMBER 1999

# THE MORAL BASIS OF POLITICAL LIBERALISM*

The exchange of essays that Jürgen Habermas and John Rawls published several years ago in this JOURNAL was a long-awaited event.[1] At last, the two greatest political philosophers of our time turned to comment directly on each other's conceptions of political legitimacy and social justice, the nature of citizenship, and the goals of political argument. We now know what Habermas and Rawls each believe, rightly or wrongly, are the strengths and weaknesses of the other's philosophy. On the whole, Habermas sought to accentuate their differences, while Rawls, though noting significant disagreements, tried to underscore the common elements in their views. I believe that in reality the two thinkers are far closer to one another than Habermas realizes, but that their proximity is not of the form that Rawls imagines. To me it seems that their fundamental point of convergence involves a common failing. Each of them, aiming at a similarly "freestanding" or "autonomous" conception of political life, misses its underlying moral foundations.

In *Political Liberalism*, Rawls[2] declares that a political conception is *freestanding*, if it looks only to the principles that should govern the political life of society. It does not present itself as applying to the political realm a comprehensive doctrine about the ends of life (PL 12). Nothing, of course, prevents such a political conception from being inte-

---

* I thank Daniel Brudney, David Estlund, Alessandro Ferrara, Rainer Forst, and Andrew Wallace, as well as the members of my seminar at the University of Chicago, for their help on earlier versions of this paper

[1] Habermas, "Reconciliation through the Public Use of Reason. Remarks on John Rawls's *Political Liberalism*," this JOURNAL, XCII, 3 (March 1995): 109-31; and Rawls, "Reply to Habermas," this JOURNAL, XCII, 3 (March 1995): 132-80. Subsequent references to these essays will be given in the text, with the page number preceded by the abbreviation 'JP'.

[2] New York: Columbia, 1993. References to this book will be given in the text, with the page number preceded by the abbreviation 'PL'.

0022-362X/99/9612/599–625

ALIPUR

grated into an encompassing moral, religious, or metaphysical view of the world. Indeed, Rawls believes that his *political liberalism*, aiming as it does to formulate liberal principles in this freestanding way, is not fully justified unless it forms a common part or "overlapping consensus" of the different visions of the good life which are prominent in modern society. But at the same time, the basic principles on which political association relies must be ones that can be framed and legitimated without appeal to those broader considerations.

I share Rawls's conviction that a liberal conception of political association should be freestanding in this sense. For liberal thought is best understood, I believe, as responding to an essential ingredient of our modern self-understanding. On fundamental issues about the meaning of life, we have come to expect that reasonable people tend naturally to disagree with one another. In early modern times, in the course of a century of religious wars, many people came to the painful realization that even with the best will in the world they would continue to differ about the nature and obligations of the true faith. Since then, this conviction has widened in scope. In a free and open discussion about the fulfilled life, the human good, or the nature of self-realization, it seems that the more we talk, the more we disagree (sometimes even with ourselves). On these matters, being reasonable—that is, thinking and conversing in good faith and applying, as best as we can, the general capacities of reason that pertain to every domain of inquiry—tends not to produce agreement but to spark controversy. Taking this experience to heart, liberal thinkers have concluded that political association should no longer undertake to express and foster a conception of the ultimate ends of human existence. Instead, it must seek its principles in a minimal morality, which reasonable people can share despite their expectably divergent religious and ethical convictions. Only so can the political world, governed as it must be by coercive rules, sanctioned by state power, come to more than just the rule of force. Only so can it enjoy the sort of transparency in which citizens recognize their political principles as the expression of their own will.

Central though it is to modern experience, the phenomenon of reasonable disagreement is not easy to explain, for it runs contrary to one of the deepest preconceptions of our philosophical tradition. Why should reason, on questions of supreme importance, work not to bring us together but to drive us apart? Historical contingencies surely play an important role. There is the great variety of life experiences created by modern Western society, with all its complex divisions of labor and rich heritage of many cultural traditions. No doubt, too, the liberal practices of toleration and public discussion have fostered this social and cultural heterogeneity, multiplying the very circumstances to which they respond. But we might also wonder whether reason does not naturally breed discord on difficult questions. To the extent that earlier societies achieved agreement about the good, or seem to have done so, does this not reflect the limits they placed on open discussion? And though the modern sciences do display widespread convergence of opinion about complex matters, might this not be because they subject the observation of nature to forms of reasoning, based on controlled experiment and measurement, which are geared precisely toward making agreement achievable?

Explaining the disharmony of reason is bound to be a speculative endeavor. Drawing out its consequences, particularly in the political realm, is more straightforward. In general, we may say, the aim of political order has always involved restraining the struggle for advantage and the violence of emotion for the sake of achieving a common good. In the past, however, this common good was typically understood as involving a comprehensive vision of human purposes, a vision which belonged to a society's most treasured knowledge, interpreted and propagated by a clerisy (often a priestly class) and sanctioned by state authority. As the modern appreciation of reasonable disagreement has grown, the fundamental tasks of political association have come to appear more complex. No longer are interest and passion alone the chief objects of concern, for the proliferation of reasonable views of the good life represents a political problem in its own right. Although people may be looking to their own interest or may be carried away by the passions of conviction when they wish to impose on others their conception of the good, they may also be acting in the disinterested belief that political life would be better if organized along its lines. It has become a distinctively modern question to ask on what basis reasonable people, thus divided, can live together nonetheless in political community.

I should point out that in referring as I have to *reasonableness*, I do not mean quite what Rawls himself does by this term. He means by it the moral commitment to seek and to abide by fair principles of cooperation, particularly in view of the extent to which people tend naturally to disagree in their comprehensive conceptions of human flourishing; thus he distinguishes the *reasonable* from the *rational*, the latter designating the intelligent pursuit of one's own ends, apart from considerations of fairness (PL 48-58). Here, as in previous writings,[3] I use the term more abstractly to mean the free and open exercise of the basic capacities of reason. *Reasonable* people according to my usage are therefore *rational*, but they may be *reasonable* in Rawls's sense as well, if they are concerned to reason about how best to be fair (though their commitment to fairness will not follow from being reasonable alone). It follows that *reasonableness* for Rawls denotes a moral response to the predicament, the profusion of disagreement about the human good, that arises from the more inclusive disposition I call being *reasonable*. In itself, the difference is only terminological. I do not deny that only in virtue of certain moral principles can liberalism give the weight it does to what I have defined as reasonable disagreement. So far am I from denying it that this moral basis forms the principal subject of my essay.

---

[3] As in my book, *The Morals of Modernity* (New York: Cambridge, 1996)

Liberalism, I have said, entails focusing on what reasonable people can still share, despite their differences about the good life. But we would badly misunderstand its nature, if we supposed its guiding principles to consist simply in whatever might turn out to be common ground among reasonable people otherwise divided by their convictions. More fundamental than the political principles on which they will agree is the very commitment to organize political life along these lines, to seek principles that can be the object of reasonable agreement. This commitment forms the moral core of liberal thought, and it embodies a principle of respect for persons.[4] Such is the main point I want to establish, as I look at the recent exchange between Habermas and Rawls and at the central question of their debate: What can it mean for political association to be founded on freestanding principles designed to abstract from the ongoing disagreements about the nature of the human good?

## I. CLASSICAL AND POLITICAL LIBERALISM

First, some historical background is necessary. We need to understand why this question should assume the greatest importance for liberal thought today.

From John Locke's time to our own, liberal thinkers have generally presented their political philosophy in terms of a full-scale individualism, urging a critical detachment toward inherited forms of belief and cultural traditions. This is not surprising. Individualism has formed a pervasive current in our culture. It grows out of basic features of modern society, particularly the market institutions of a capitalist economy. It has also drawn encouragement from the modern experience of reasonable disagreement, as people have concluded that they must therefore determine on their own their vision of the human good; and that conclusion has seemed to define the basis on which they must understand the terms of their collective life. Thus, the classical liberalism of Locke, Immanuel Kant, and J. S. Mill argued that the principles of political life should forego appeal to ideals of the good because thereby they express the individualist spirit which ought to shape the whole of our life. Classical liberals differed, of course, in important ways concerning the precise nature of the individualism they embraced. But they agreed on the fundamental idea that our allegiance to any substantial view of the good life—to any concrete way of life involving a specific structure of purposes, significances, and activities (such as a life shaped by certain cultural traditions, or devoted to a particular religion)—should always be a contingent one, revisable on reflection. Such forms of life can be truly valuable, they believed, only if we understand them as ones we choose, or would choose, from a position of critical reflection. Most importantly, they were at one in defending their political principles within the framework of this general philosophy of individualism. Our status as political subjects or citizens should be independent of whatever view of the human good we affirm, because in that way political principles respect—as Locke, Kant, and Mill would have said—the fallibilist, autonomous, or experimental attitude which we as persons should maintain at the deepest level of our self-understanding.

---

[4] Despite its philosophical salience, this moral commitment does not, I believe, provide a sufficient basis for a liberal political order. It cannot explain why people who hold this commitment should think of themselves as *a people*, bound together in a common political destiny which distinguishes them from others. This element in their self-understanding arises from a shared historical experience, particularly the memory of past conflicts which were fueled by the effort to impose faiths or other comprehensive ideals on one another. For some more on this point, see my *Morals of Modernity*, pp. 141-44.

Things, however, have not stayed so simple. The individualist view of life has itself become an object of reasonable disagreement. Especially in the wake of the Romantic movement, there has arisen a new sense of the significance of belonging, an appreciation of tradition to which the premium that individualism places upon critical reflection appears to embody a kind of moral blindness.[5] In reality, a reflective attitude toward received forms of life is one value among others. To give it supreme authority may therefore block recognition of much else that is also of value. Thus it has been held that we can share in the good that some ways of life offer, only if we do not think of our allegiance to them as elective, as a matter of decision, but regard it instead as constitutive of what we hold to be valuable, as rooted in a feeling of belonging. The importance of common customs, ties of place and language, and religious faith can lie in their shaping the very understanding of value on the basis of which we make the choices we do. At the very least, is it not difficult to imagine our deepest moral commitments as ones that, from a critical distance, we choose or would choose to affirm? To stand back in reflection from them would seem to require stripping ourselves of the resources necessary to guide any sort of moral evaluation. Ought they not to be understood, therefore, as the inherited basis of deliberation, rather than as objects of choice themselves?

---

[5] For a more detailed account of this Romantic theme, see my book, *The Romantic Legacy* (New York: Columbia, 1996), chapter 2, as well as *The Morals of Modernity*, pp. 127-34.

The praise of belonging, though often presenting its critique of individualism as an opposition to the Enlightenment, does not involve a reversion to premodern forms of thought. The forms of life with which it encourages identification were seldom tradition-minded themselves, seeking instead legitimation in some transcendent source, such as the voice of Reason or the will of God. Paradoxical as it may sound, traditionalism is a modern innovation, and for that reason we can expect it to prove an abiding feature of our culture. Today, despite the enormous influence they continue to wield, individualist modes of thought have turned eminently controversial. No doubt they have always been contested, but now their philosophical difficulties are open to view. However we settle to our own satisfaction the merits of individualism and the sense of tradition as conceptions of the good and the right, we cannot deny that on this question, reasonable people continue to disagree.

In this regard, then, liberalism faces a new challenge. Should it keep its classical commitment to an individualist view of life? Or, taking instead its bearings from that modern experience of reasonable disagreement that was in the beginning its guide, should it seek a reformulation broad enough to appeal to those who are otherwise divided by this controversy? The second path is the one pursued by political liberalism, as Rawls and I understand it. The aim is to fix the principles of political association in terms independent, not just of religious convictions and substantive notions of the good life, but of comprehensive moral conceptions, too, to the extent that espousing individualist ideals or their rejection, they have become an object of reasonable disagreement in their own right. So understood, political liberalism does not represent a radical departure from the motivations of its classical forebears. The underlying continuity is the focus on the phenomenon of reasonable disagreement, and the differences stem from experience, as we have learned how pervasive this phenomenon actually is. There may not therefore be much of a puzzle as to why political liberalism did not emerge earlier. It took the Romantic critique of individualism and its absorption into our culture to make plain how controversial reasonable people can find the overall moral vision to which classical liberalism has appealed.[8]

## II. POLITICAL LEGITIMACY AND MORAL RESPECT

Historical experience provides the impetus for liberalism to shed its individualist philosophy. Nonetheless, in assuming the form of what Rawls calls a freestanding political conception, liberal doctrine is heeding more than just the widening scope of reasonable disagreement. It is also drawing upon certain moral convictions, which imply that this is the proper route to take. After all, we may intelligibly ask why liberalism's response to this controversy should be a reformulation of its principles. Why should liberalism become *political*, in the sense that Rawls and I intend? Why should liberal thinkers not instead dig in their heels and, observing correctly that no political conception can accommodate every point of view, maintain that liberalism stands or falls with a general commitment to individualism? The answer must be that the essential convictions of liberal thought lie at a more fundamental moral level. So we need to bring out what these convictions are.

Rawls remarks that liberalism seeks the form of a freestanding political ideal, because it "applies the principle of toleration to philosophy itself." By this he means that, because its goal is a conception of justice to which reasonable people can agree, this conception must "be, as far as possible, independent of the opposing and conflicting philosophical and religious doctrines that citizens affirm" (PL 9-10). In other words, the reason why liberalism must abandon its classical appeal to an overall individualism and become a strictly political doctrine lies in the idea that basic political principles should be acceptable to those whom they are to bind. This idea Rawls has come to call the *liberal principle of legitimacy*. "Our exercise of political power is fully proper," he explains, "only when it is exercised in accordance with a constitution the essentials of which all citizens as free and equal may reasonably be expected to endorse in the light of principles and ideals acceptable to their common human reason" (PL

---

[8] In giving such weight to this historical perspective, I may differ from Rawls, who finds puzzling the late development of political liberalism. See his "Reply to Habermas," JP 133, footnote 1.

137).[7] This principle, I agree, reflects the abiding moral heart of liberal thought.[8] But I do not think that Rawls has brought out as distinctly as he might the nature of this moral foundation and the precise position it occupies in political liberalism.

To see better what the liberal principle of legitimacy involves, let us ask why we believe, if we do, that the fundamental terms of political life should be the object of reasonable agreement. Answering this question will prove essential, in fact, to determining what reasonable agreement should mean in this context. Do we accept the liberal principle of legitimacy, because we think that in general people are bound only by moral rules they could not reasonably reject? Whether correct or not, this view of morality is far too controversial for the present purpose. As an explanation of the grounds of moral obligation,[9] it risks circularity: Do not the reasons we have to accept the fundamental rules of morality involve the fact that it is simply wrong not to observe them? Moreover, in this form the view seems difficult to detach from the broader moral ideal of individual autonomy or self-determination toward which political liberalism must strive to be neutral. And if the validity of moral rules is instead supposed, not to be explained by reasonable agreement, but only to imply it,[10] the view still remains too controversial. For there is more than a little sense in holding that a person stands under the basic obligations of morality even if (perhaps through bad living) he has come to see the world in a way that can only be indifferent to them. In any case, general conceptions of moral obligation are not, at least directly, the source of our conviction that political principles must be able to meet with the reasonable agreement of the citizens they are to govern.

That conviction reflects instead the distinctive feature of political principles which sets them off from the other moral rules to which we may believe people are subject. Moral principles fall into two groups. With some we believe people can be rightly forced to comply, but others we do not regard as valid objects of enforcement, whatever disapproval or even outrage we may feel when they are violated. The first group alone has the status of political principles. For an association is political insofar as it relies upon the supposedly legitimate use of force to secure compliance with its rules. It is this coercive character of political principles which we have in mind, when we hold with the assurance that we do, and whatever our allegiance to the general ideal of judging people only by rules they could find acceptable, that such principles must be the object of reasonable agreement. Our belief is that only so can the use of force to implement these principles be justified. This, incidentally, is why political liberalism, though claiming to be a strictly political conception, does not imagine that the political domain forms a prepackaged sector of society, inherently distinct from the other areas of social life.[11] It is we who decide what will count as *political*, depending on what elements of social behavior we think should be subject to coercion.

---

[7] For other references to this principle, see PL 139–40, 143, 217, 225–26.

[8] Following Rawls, I also agree that this principle is meant to govern chiefly the choice of basic, constitutional principles. These principles will themselves allow for many decisions to be made according to less demanding rules, such as majority voting, in which indeed appeals to controversial ideals of the good may sometimes be appropriate.

[9] Such a conception of morality is presented by T. M. Scanlon in *What We Owe to Each Other* (Cambridge: Harvard, 1998).

[10] This appears to be the basis on which Gerald Gaus argues that political principles require public justification—*Justificatory Liberalism* (New York: Oxford, 1996), pp. 121ff.

With all these remarks Rawls himself would concur, for at one point he notes that the liberal idea of legitimacy rests on the fact that political power is coercive power (PL 139). But our analysis of this core liberal commitment must go deeper still. We need to make clear why it is that the validity of coercive principles should depend upon reasonable agreement.

I believe that the source of this conviction is a principle of *respect for persons*. Let me explain.[12] Observe first that the use or threat of force cannot be deemed wrong in itself, for then political association would be impossible. What we must regard as improper is rather to seek compliance by force alone, without requiring reasonable agreement about the rules to be enforced. For consider the basic fact that persons are beings capable of thinking and acting on the basis of reasons. If we try to bring about conformity to a rule of conduct solely by the threat of force, we shall be treating persons merely as means, as objects of coercion, and not also as ends, engaging directly their distinctive capacity as persons. True, they cannot be moved by threats except by seeing that they have good reason to fear what we may do. But we shall be appealing to their ability to act on reasons simply in order to achieve the goals of compliance—the establishment of public order, perhaps also the reformation of people's character. We shall not be engaging their distinctive capacity as persons in the same way we engage our own, making the acceptability of the principle depend on their reason just as we believe it draws upon our own. Thus, to respect another person as an end is to require that coercive or political principles be as justifiable to that person as they presumably are to us. This is certainly not the only sense we can give to the rich moral notion of respect.[13] But it is the one which liberals must regard as relevant from a political point of view. (In this regard, the extent to which it avoids the peculiarities of the Kantian conception of respect is an advantage.)

---

[11] Habermas makes this criticism of Rawls's political liberalism, wrongly it seems to me, in "Reconciliation through the Public Use of Reason," JP 129.

[12] Here, I rely on the argument laid out in *The Morals of Modernity*, pp. 136-41.

## III. THE MORAL FOOTING OF A FREESTANDING POLITICS

Liberalism, formulated as a strictly political doctrine, rests therefore on this moral foundation. It forms a freestanding conception in regard to comprehensive moral visions of the good life, but it cannot coherently claim to be freestanding with respect to morality altogether. In particular, we would be wrong to suppose that the moral principle of respect for persons has the political significance it does because reasonable people share a commitment to it. On the contrary, the idea of respect is what directs us to seek the principles of our political life in the area of reasonable agreement. Respect for persons lies at the heart of political liberalism, not because looking for common ground we find it there, but because it is what impels us to look for common ground at all.

These reflections lead, I believe, to a very important result. As citizens of a political association organized in accord with the liberal principle of legitimacy, we cannot regard the norm of respect as having the same sort of validity as the constitutional principles by which we live. Those principles are legitimate in virtue of being the object of reasonable agreement, and thus their authority is strictly political, deriving from our collective will as citizens. Or at least so they are understood for political purposes (for when we refer to our global visions of the good, we may also attribute to them an independent status). But even in our capacity simply as citizens—relying solely on the commitments we share as members of a political community—we cannot take this view of the principle of respect. It must instead be understood as having more than just political authority. We must consider respect for persons as a norm binding on us independent of our will as citizens, enjoying a moral authority that we have not fashioned ourselves. For only so can we make sense of why we are moved to give our political life the consensual shape it has.

---

[13] Thus, as William Galston points out—*Liberal Purposes* (New York: Cambridge, 1991), p. 109—we respect a person in a different sense when explaining to him fully our reasons for the principle by which we judge his conduct, whether or not he can appreciate those reasons. My point is not to settle what respect "really" means, for it really means a great many different things. My concern is with what it ought to mean when figuring at the foundations of liberal thought. Also, I agree with Rawls—*A Theory of Justice* (Cambridge: Harvard, 1971), pp. 585-86—that the relevant notion of respect is best circumscribed by referring to characteristic liberal principles (such as the liberal idea of legitimacy). But our aim ought then to be understood as determining the sense of respect for persons on which those principles rest.

It is not clear to me that Rawls understands his own thought in this way. To be sure, he thinks of his political liberalism as a "moral conception" (PL 11n). In holding that liberal principles should be formulated as a strictly political conception, he does not mean to suggest any contrast between the political and the moral, as though liberal doctrine did not form a normative conception, as it manifestly does, consisting in certain ideals, standards, and values. But would he be willing to admit that, as citizens reasoning from the standpoint of this political conception, we must acknowledge a moral authority higher than the political principles we give ourselves?

No unambiguous answer is available. About the liberal principle of legitimacy, Rawls says that it has the same basis as his two principles of justice: it would be chosen in what he calls the "original position," being indeed bound up in those very two principles (PL 137n, 225-26). In other words, Rawls seems to regard that principle of legitimacy as one whose validity, at least from the political point of view, depends on our collective will as citizens, whereas I of course have been arguing for just the opposite conclusion. Yet Rawls also notes that the original position, which is but a device for representing the basis on which we understand ourselves as choosing or imposing on ourselves political principles, incorporates certain values in the conditions under which such a choice is to be imagined taking place, values which thus are not themselves the object of choice (for example, PL 103). Although we do not, Rawls stipulates, imagine the parties choosing in the original position as endowed with any moral sensibility (they are merely *rational*, engaged in the efficient pursuit of their ends), the fact that we place certain conditions on their choice—they are to be supposed ignorant of their own talents or conception of the good life, for example—reflects a moral commitment of our own, namely, a commitment to what he calls reasonableness, the readiness to seek fair principles of cooperation (PL 305). Political principles apparently have then a moral basis that we cannot conceive as rooted like them in our political will. Moreover, Rawls describes this moral commitment of reasonableness as embodying the resolve to propose political norms justifiable to all (PL 49, 50); and this brings us right back to the liberal principle of legitimacy. That principle cannot therefore have the same status as the two principles of justice; and, as it expresses in effect the idea of respect for persons, Rawls would apparently agree after all that this idea must have a moral authority for citizens that is independent of their political will.

Yet Rawls never makes this point explicitly. On the contrary, his stated view is that because of its freestanding character, political liberalism is "doctrinally autonomous" (PL 98-99). At least in our role as citizens (if not in our comprehensive views of the world), we are to regard our political principles as deriving their validity from our political will. These principles cannot accordingly be understood, he holds, as "moral requirements externally imposed" (PL 98). This phrase can have various meanings, however. If it means that basic political principles are not to be imposed upon a citizenry by some external agency—by an enlightened monarch, for example—then I agree.[14] But if instead it means that citizens should not regard their political principles as drawing upon moral requirements whose validity is external to their collective will, then I believe Rawls goes wrong. Political liberalism makes sense only in the light of an acknowledgement of such a higher moral authority.

In fact, the idea of respect plays this foundational role in several ways. As we have seen, it forms the basis for believing that political principles should be the object of reasonable agreement. But, in addition, it serves to define the very nature of the agreement to be sought. To explain this point, I begin by observing that consensus in this case is clearly a normative notion if only because it refers to *reasonable* instead of simply *actual* agreement. Political life is to be based on principles that citizens, despite their various moral, religious, and metaphysical beliefs, can see reason to accept, exercising (as I have said) the basic capacities of reason applicable to any domain of inquiry.[15] But reasonableness (so understood) is not the sole normative criterion defining the sort of consensus in question. Because the idea of respect directs us to look for common ground in the first place, it should be understood as a further condition which acceptable principles must satisfy. The terms of political association are to be judged by reference to what citizens would accept, were they reasonable and committed to the principle of equal respect for persons. The notion of agreement to which political liberalism appeals is therefore an idealization. It comes into play only within the bounds set by these two norms, the one epistemic, the other moral.

---

[14] In this connection, see Rawls, "Reply to Habermas," JP 162-63.

[15] The reasonable so understood does not entail uncritical deference to common sense and to the formal fallacies of reasoning which it may harbor or to the refusal to admit the existence of reasonable disagreement about the good which it may show. Contrary to Gaus—*Justificatory Liberalism*, pp. 3-5, 131-36—political liberalism as I conceive it (or, I believe, Rawls as well) is not *populist* in this sense.

This conclusion is, again, one which Rawls should be willing to admit. His liberal principle of legitimacy ties consensus to precisely these two assumptions when referring to what citizens in the light of "their common human reason" may be "reasonably expected to endorse." (Recall that his notion of the reasonable embodies the principle of respect.) Yet he is not as clear as he should be about the moral foundations of this idea of consensus.

## IV WHAT HABERMAS AND RAWLS SHARE

Habermas, too, aims to develop an *autonomous* conception of the guiding principles of modern democracy. This ambition inspires the defense of radical democracy in his book, *Faktizität und Geltung: Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaates.*[16] The democratic ideal as Habermas understands it has nothing to do with the communitarian dream that our political life might once again draw its bearings from a common vision of the human good. In this respect, it shares the outlook of political liberalism. But Habermas also opposes it to what he calls the liberal idea that individual rights set limits to the exercise of democratic self-rule. Liberal thinkers err, he believes, in thus holding political association accountable to moral norms supposedly given in advance. For they thereby fail to heed the intellectual conditions of modern times. In an age where religious and metaphysical worldviews have lost their authority, we can have reason to consider ourselves subject to political principles only if we are able at the same time to see ourselves as the authors of these principles:

---

[16] Suhrkamp: Frankfurt, 1992, p. 13; English translation by William Rehg: *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy* (Cambridge: MIT, 1996), p. xlii. Subsequent references to this work will be given in the text, with 'FG' preceding the page in the German original and 'BFN' the corresponding page in the translation. In quotations, I have changed the English translation as I saw fit.

> Without religious or metaphysical support, the coercive law designed for legal behavior can preserve its socially integrative force only insofar as the *addressees* (*Adressaten*) of legal norms may at the same time understand themselves, in their collectivity, as the rational *authors* (*Urheber*) of those norms (BFN 33; FG 51-52).

In Habermas's view, this kind of collective autonomy (to be distinguished, obviously, from the ideal of individual autonomy, though the two are structurally similar) constitutes the source of the basic norms shaping political association. We miss the true character of the principles of modern democracy, he claims, if we adhere to the "classical" (that is, premodern) notion of a normative hierarchy whose pinnacle is occupied by independent moral principles to which political association must defer (FG 137; BFN 106, 449). Because liberal thought, subordinating democratic self-rule to individual rights, continues to rely on this hierarchical model, Habermas believes that we must move beyond a liberal conception of political life. We must be resolutely modern, recognizing that political principles have their basis only in the autonomous will of citizens who are the full authors of the rules that bind them.

Habermas often describes his position as one in which fundamental individual rights, no longer serving to check democratic self-rule, turn out to be *co-original* (*gleichursprünglich*) with it. In other words, the modern and ancient ideals of liberty, as those two principles have come to be called, are not so open to conflict as commonly supposed.[17] Sometimes Habermas's idea seems to be that the two mutually support one another. Just as self-government can serve to protect individual rights, so these rights themselves provide the necessary means for the exercise of popular sovereignty—the right to free expression, for example, making available the wide range of information and points of view that permit wise political decisions (FG 155, 161; BFN 122, 127-28).

But, in reaching for the idea of co-originality, Habermas has clearly a deeper point in mind as well. It is that the two principles have a common origin, and that source is, he believes, the autonomy by which a community gives shape to its political life. The result is that, despite his talk of mutual implication, Habermas unmistakably makes democratic self-rule a principle prior in status to that of individual rights. The claim that rights and self-rule are co-original is misleading, in light of what he takes to be the only plausible origin of political norms in the modern age. In reality, popular sovereignty functions for Habermas as the ultimate basis on which our political life should be organized, and so as the true source of individual rights, their ultimate justification lying in the way they embody the principle of self-rule (cf. FG 134-35; BFN 104). The reason is plain: only so can we escape letting the principles of political association draw upon moral values having an independent validity. (I come back to the details of Habermas's argument in section VII.)

---

[17] See Habermas, *Die Einbeziehung des Anderen* (Frankfurt: Suhrkamp, 1996), p. 298; English translation by Ciaran Cronin and Pablo De Greif: *The Inclusion of the Other: Studies in Political Theory* (Cambridge: MIT, 1991).

A striking similarity thus emerges between Habermas's conception of autonomy and Rawls's demand that political principles be freestanding. Indeed, as I have remarked, Rawls himself calls a freestanding political conception "doctrinally autonomous" and explains this term in much the same way. A political conception is autonomous, he writes, if "the political values of justice and public reason (expressed by their principles) are not simply represented as moral requirements externally imposed" (PL 98). In their companion essays in this JOURNAL, both philosophers emphasize this point of convergence (JP 127, 150).

But it is also obvious that Habermas's idea of political autonomy is the very notion that I have criticized in my discussion of Rawls. Political association, so we have seen him claim, must not be understood as drawing its principles from a moral source superior to its collective will. In his hands, this conception fares no better than it does in Rawls's. Habermas, too, misses the moral basis that supports the democratic ideal of self-rule. To make clear where precisely his mistake lies, I shall first look at some disagreements between the two philosophers. In their debate in this JOURNAL, they both noted important differences in their handling of this common idea. Rawls rejected the manner in which Habermas arrives at his conception of political autonomy. Habermas, in turn, charged that Rawls strays from a consistent development of such a conception. Understanding these points of divergence will put us in a better position to see why, contrary to them both, modern democracy rests upon independent moral foundations.

## V. METAPHYSICS AND POLITICS

As Rawls observes (JP 135-38), political autonomy in Habermas's eyes stems from a broader philosophical position described as *postmetaphysical*. In our time, Habermas declares, metaphysical theories asserting the existence of entities that are neither physical nor psychological in nature ("ideal entities" he calls them) as well as religious conceptions that see the world as the work of God and the vehicle of providence have lost their plausibility. His point is not simply that they no longer enjoy the society-wide authority they once possessed, having become the object of apparently irresoluble disagreement. For Habermas, the age of metaphysical and religious worldviews is over in the deeper sense that, properly speaking, rational belief in their truth is no longer possible. Reason itself, he believes, now presents itself as essentially finite, fallible, procedural, and oriented toward intersubjective agreement. It consists in thinking or acting on the basis of reasons we take to be valid, not because we imagine them guaranteed by an objective order of the world, but because we suppose that they would command the assent of others under appropriately ideal conditions. We are to locate the norms of reason, no longer in an ideal realm independent of us, but instead in the idealizations we necessarily make in speaking with one another.[18] A battery of philosophical arguments, assembled under the title of *discourse theory*, aims to bring out the necessity of this postmetaphysical shift in our conception of reason and to explain the idealized form of discussion (*Diskurs*) to which reason thus refers.

Habermas claims that it is in virtue of this general standpoint that the idea of political autonomy recommends itself. We are to establish the terms of political association by applying this general account of reason to the fundamental task of politics, which is to determine the scope of the rule of law. Thus, once reason can invoke no higher authority than agreement under ideal conditions, the principles of conduct that are to bind us must be such that we can see ourselves to be their source. This is the meaning of the passage I began by quoting in section IV.

Rawls takes his distance from this approach, and rightly so. Habermas's argument has the significant disadvantage that it relies on a comprehensive vision of our place in the world and the nature of reason's authority. His postmetaphysical point of view shares, as Rawls insightfully notes, the ambition of G. W. F. Hegel's logic, providing "a philosophical analysis of the presuppositions of rational discourse...which includes within itself all the allegedly substantial elements of religious and metaphysical doctrines" (JP 137). Such doctrines will scarcely recognize, much less accept, the form into which they are thus translated, for it is one in which, as Habermas himself remarks, "worldviews are measured more by the authenticity of the life styles they shape than by the truth of the statements they admit" (JP 126). This idea of postmetaphysical reason can be expected therefore to provoke controversies of precisely the sort that we must look beyond, so Rawls believes (and I agree), if we are to discover acceptable terms of political association. After all, the notion of ideal discussion to which Habermas appeals is not self-explanatory, and the transcendent interpretation he favors, defining ideal conditions as he does by reference, not to the ("local") standards of belief we currently espouse, but to the best standards we ever could have (FG 30-31, 36, 202, 566; BFN 15, 20, 163-64), is neither the only possibility nor one likely to strike many people as anything short of metaphysical.

---

[18] There is a succinct and lucid account of this line of thought in FG 24-37; BFN 9-21.

In truth, the vitality or obsolescence of metaphysical and religious worldviews is an issue on which reasonable people tend naturally to disagree. It is fair to observe that today none of these conceptions can prove authoritative for society as a whole. Yet to go further and say, as Habermas does, that they can no longer figure as objects of rational belief, but only as life styles, means adopting a point of view that itself is party to these disputes. Habermas has misidentified the feature of modern experience that is crucially relevant to the basis of political association. The decisive element is not the waning of metaphysical and religious worldviews (though that has occurred). It is instead the recognition that such worldviews, as well as the recurrent postmetaphysical efforts to do without them, are an enduring object of reasonable disagreement.

As a matter of fact, Habermas's idea of postmetaphysical reason seems to me to go wrong in two important respects.[19] First, I believe that our conception of the world must have room for ideal entities, for only if reasons exist (a reason being itself neither physical nor psychological in character) can there be such a thing as normative knowledge—knowledge of how we ought to act, but also more fundamentally knowledge of how we ought to think. Second, I do not think that the notion of ideal discussion forms a substantial part of the theory of justified belief. We may say, if we wish, that justified beliefs are ones to which we assume that all would agree in an "ideal" discussion. But we can have a grasp of such ideal conditions only if we suppose them to embody our best standards of warranted belief, and so we might as well say, more directly, that the beliefs are justified if they satisfy these standards.[39] In Habermas's account of justified belief, it is true, ideal discussion plays a role that resists elimination of this sort, since it is assumed to transcend our current standards. But in consequence it loses any discernible content.

---

[19] For more on these two points, see my *Morals of Modernity*, chapter 5 and pp. 206-10

Yet I should emphasize that the main point in the present context is not whether Habermas's conception of reason is correct, but whether it can serve as an appropriate basis for establishing the terms of our political life. These two questions must be distinguished, once we are convinced that political association turns crucially on finding principles reasonable people can accept despite their disagreements about fundamental matters of life and value. In seeking a solution to this political problem, we cannot call upon all that we ourselves may reasonably believe to be true about such matters. Habermas gives little attention to the difference between these two perspectives. That no doubt is why he runs them together. The result is an account of the foundations of political life which, whatever its interest, and whatever the ultimate validity of the conception of reason on which it relies, is destined to be an object of reasonable disagreement. In this regard, Habermas's political theory falls short of what we need.

## VI. HABERMAS'S IDEAL OF POLITICAL AUTONOMY

Habermas himself locates the crucial difference in their accounts of political autonomy elsewhere. In his view, Rawls fails to adhere consistently to this principle in working out his theory of justice, reverting to the liberal assumption that there exist individual rights prior to democratic self-rule and defining its scope. This backsliding, he charges in the exchange in this JOURNAL, arises from the "two-stage character of his theory" (JP 128). While Rawls's idea of the original position, in which rational parties choose principles of justice, captures something like collective autonomy, it yields a conception of the just society in which "liberal rights...constrain democratic self-legislation" (JP 128-29). Citizens cannot then "reignite the radical democratic embers of the original position in the civic life of their society" (JP 128).

Behind this objection there may well lie, as Rawls suspects (JP 160), some sympathy with the quixotic idea (once famously voiced by Thomas Jefferson) that every generation should give itself its own constitution, unfettered by the past decisions of those now dead. But I believe that the gist of Habermas's complaint has little to do with a call for permanent revolution. It concerns instead the understanding we should have of the basis of our political association. His view is that citizens must be able to see all their political principles, even those establishing individual rights, as rooted in their autonomous political will. Such a self-understanding is blocked by a conception that gives rights, as he believes Rawls's theory does, a status prior to the democratic process. This is the point at which liberalism and Habermas's own theory of radical democracy supposedly part ways.

---

[39] There is a similar criticism in Albrecht Wellmer, *Ethik und Dialog* (Frankfurt: Suhrkamp, 1986), pp. 70-80; English translation by David Midgley included in *The Persistence of Modernity: Essays on Aesthetics, Ethics, and Postmodernism* (Cambridge: MIT, 1991).

What is the force of Habermas's criticism? I am inclined to think that he terms *inconsistency* what I have called *ambiguity*. Habermas objects that Rawls fails to carry through consistently an autonomous or freestanding conception of political life. But perhaps another way to put the point is to say, as I have argued, that Rawls's idea of a freestanding political conception is ambiguous. Sometimes he implies and sometimes denies that his strictly political doctrine rests on the basis of a moral principle of equal respect. To the extent that he admits this truth, his understanding of a freestanding political conception does not really correspond to what Habermas calls autonomy, which excludes this sort of normative hierarchy. In any case, it is clear that Habermas and I take opposite views of these tensions in Rawls's thought.

But there is a further complication. Habermas's own political theory does not hold consistently to the conditions he assigns to the notion of autonomy. In fact, his conception of political autonomy makes sense only on the assumption that self-rule rests on an underlying, moral norm of respect. Nor is that astonishing. For only on such a moral basis can popular sovereignty take the form that we prize as modern democracy. Moreover, this moral foundation involves, contrary to Habermas's stated intention, an idea of individual rights that precedes and defines the exercise of self-rule. In the end, therefore, the sharp opposition he constructs between liberalism and his own conception of radical democracy, and thus the criticism he makes of Rawls, lose their raison d'être. Habermas, I believe, misunderstands himself. It is to developing these points that I now turn.

To begin, let us examine more closely the way Habermas understands political autonomy. In explaining how the autonomy characteristic of modern democracy must proceed without relying upon any antecedent moral norms, Habermas appeals to what he calls *discourse principle D*:

> Just those action norms are valid to which all possibly affected persons could agree as participants in rational discourses (*Diskursen*) (BFN 107; FG 138).

Modern democracy has its basis in the collective autonomy of its citizens, Habermas declares, because in determining the principles by which they will govern themselves, they defer to no pregiven norms other than just this principle *D*. "Nothing is given prior to (*vorgegeben*) the citizens' practice of self-determination (*Selbstbestimmungspraxis*) other than the discourse principle" (BFN 127-28; FG 161). In contrast to liberal theories, which appeal to antecedent moral norms (in the form of individual rights) to limit the scope of self-rule, radical democracy, as Habermas conceives it, looks to no higher authority than its own self-legislative activity as governed by principle *D*. Rights guaranteeing individual freedoms (such as speech and assembly) have their basis, he claims, in the application of the principle to the very form of a legal system, whereas the rights of political participation serve to make possible the institutionalization of this sort of self-rule (FG 154-62; BFN 121-28).[21]

My concern lies, not with the details of this "logical genesis of rights," but with the status of *D* itself. It is crucial for Habermas's conception of radical democracy, as he makes clear (FG 138; BFN 107), that principle *D* possess no moral content of its own.[22] Morality and democracy, he holds, involve distinct and equally basic applications of principle *D*, the difference consisting in whether the principle is to shape the interactions between human beings as such or is to take a legal form binding for a specific political community. Certainly, there is a significant overlap between the norms thus generated in these two domains. But in Habermas's view, this does not change the fact that the terms of political life have their justification solely in the autonomous will of the citizens, as expressed according to principle *D*, and not in any moral principles that citizens must recognize as setting limits to their will.

This is the point, however, where Habermas displays an incomplete understanding of his very devotion to democratic self-rule. Observe that popular sovereignty itself may take different forms, depending on how the popular will is supposedly determined and given authoritative expression. Even modern dictatorships, since unlike premodern monarchies they disclaim dynastic or divine legitimacy, are a phenomenon of the modern democratic age: they too draw their authority from the will of the people, as they choose to understand it.[23] When conceived as heeding a principle such as *D*, popular sovereignty takes on a form distinctive precisely by virtue of its moral character, for all that Habermas may protest to the contrary. *D*, he admits, possesses "normative" content, even if it is not a moral principle (FG 138; BFN 107). That is surely a feeble distinction. *D* has moral content, and we can bring it to light by asking the simple question: Why should we believe, as this principle requires, that norms of action must be rationally acceptable to all whom they are to bind? Whence arises the authority of principle *D* itself?

---

[21] See also Habermas, *Die Einbeziehung des Anderen*, pp. 299-301.

[22] As Habermas observes (FG 140; BFN 108), this thesis of *Faktizität und Geltung* represents a departure from his earlier writings. I think it is a wrong move.

## VII. DEMOCRACY AND LIBERALISM

Habermas himself never provides an adequate answer to this question. He suggests that such a principle is inevitable, once we realize that in our age metaphysical and religious conceptions, characteristically inclined to invoke principles of conduct of which we are not the source, have lost their rational warrant. But this response will not work on a variety of counts. As I have already indicated, it is rash to imagine that reason today stands uncontroversially opposed to belief in the truth of metaphysical or religious worldviews. A more accurate diagnosis of our situation is that such visions can no longer enjoy reasonable agreement within society as a whole. We might well take this fact to entail that metaphysical and religious conceptions should not figure among the principles that bind us together politically. But note that such a conclusion is valid, only if we already accept some principle such as *D*; it cannot serve to justify that principle. Observe further that, if we refuse a political role to metaphysical and religious worldviews on some other grounds—because, for example, we believe, as Habermas does, that they can no longer be rationally held to be true, we shall not thereby have reason to adopt principle *D*. Why should we not suppose that without appeal to such ultimate truths the problem of political association is insoluble? Or why should we not organize our political life around some principle other than *D*, such as the maximization of the general welfare?

Sometimes Habermas claims that the principle enjoys a privileged status because it inheres in the very idea of discussing with others the norms to which our actions should conform, that idea being the only common resource we can count on in an age of controversy.[24] This suggestion continues Habermas's longstanding interest in basing morality on "universal pragmatics." I doubt, however, that so much can be gotten for so little. *D* is too substantial a principle, and precisely because of its significant moral content, to be implicit in the mere idea of practical discussion.

---

[23] This point was rightly made, however disingenuously for his own purposes, by Carl Schmitt, *The Crisis of Parliamentary Democracy* (Cambridge: MIT, 1985), p. 28 (German original: *Die geistesgeschichtliche Lage des heutigen Parlamentarismus* (Berlin: Duncker and Humblot, 1923 and 1926)).

A good way to see this is to reflect that, taken as a general principle applicable to all domains of life, *D* is not even obviously correct. Of course, the very meaning of *D* is obscure so long as the notion of "rational discourse" is left undefined, and Habermas's own wish to understand that notion in abstraction from any specific standards of belief does not help. But however we explicate the idea of rational agreement, principle *D* in its general form faces the sort of objections I mentioned in section II. Thus, many people believe indeed that we should not judge others by rules that on due consideration they would not themselves accept. But many, too, believe the opposite. They are convinced that certain fundamental norms of conduct, possessing an objective validity, can serve as a basis for judging others, whether or not they meet with agreement. I myself subscribe to such a view. But what should be clear is that the general validity of a principle such as *D* is a matter on which reasonable people disagree. Equally plain is the fact that this disagreement turns on different moral convictions about the conditions under which we may judge others morally and no doubt, too, on different appreciations of the moral ideal of individual autonomy. This suffices to show that contrary to Habermas, *D*, taken as a general principle, has a moral content and a controversial one at that.

But let us go further and look at *D* in the political role it is supposed to play. Recall a point made earlier: norms of conduct are of two kinds. There are, on the one hand, norms that we invoke solely to judge the conduct of others, and, on the other, norms that we believe should be backed up by the threat and use of force to ensure compliance. So let us consider principle *D* as applied simply to norms of the second kind, that is, to *political* principles. Now its complexion is rather different. It need not be rejected by someone who believes that we may rightly judge others in the light of norms that they themselves would not accept. For it requires only that the principles of our political life, being coercive, be rationally transparent to those whom they are to bind. In this limited, political form, *D* thus amounts to the conviction that Rawls calls the *liberal principle of legitimacy*. As a result, however, it cannot be right for Habermas to regard *D* as independent of any antecedent moral commitments. For, as I have argued, the liberal principle of legitimacy embodies a moral principle of respect for persons. This moral foundation is what gives *D* the political authority it enjoys. Thus, Habermas's conception of democracy fails to be autonomous in the absolute way he desires. If we believe our political life should be organized by some principle such as *D*, that is only because we embrace the moral principle of equal respect for persons.

---

[24] See, for example, most recently Habermas, *Die Einbeziehung des Anderen*, pp. 58-59.

The sort of normative hierarchy that Habermas dismisses as premodern shapes in fact his own thinking. Self-government, so far as it admits only those political principles which can be the object of reasonable agreement, rests on an independent moral basis. We cannot regard respect for one another as just one more political principle among others whose validity derives from our political will. On the contrary, this principle justifies the ideal of self-government and defines for us what it means. Reasonable agreement must in this context be understood (as I argued in section III) as circumscribed by the moral principle of respect. Comprehensive moral views of the ends of life being eminently controversial, our political life cannot seek its basis in them. About this point Habermas is right. But we cannot suppose, as he does, that our political life should be similarly freestanding or autonomous with respect to morality altogether. Citizens can understand themselves as the collective author of their political principles, only if they see themselves as already bound by the claims of respect for persons.

This moral principle refers, moreover, to an individual right, even if one more fundamental than the political rights that usually figure among explicit constitutional guarantees. Every person has the right, it claims, to be bound only by political principles whose justification he can rationally accept. As a result, Habermas's position comes down to one in which, contrary to his own account, an individual right sets limits to democratic self-rule, determining as it does the sort of expression of the popular will that shall count as *democratic*. The familiar constitutional rights of free-expression, property, and political participation, though no doubt serving to promote the goal of democratic self-rule, also have an independent rationale. They draw upon that most fundamental of individual rights, which is the right to equal respect.

In the Postscript to the English translation of *Faktizität und Geltung*, Habermas has replied to an earlier version of this criticism.[8] To understand the normative basis of modern democracy, he argues, we must distinguish the *horizontal sociation of citizens*, in which they constitute their collective will and in so doing mutually accord rights to one another, from the subsequent step in which they set up rights to protect individuals from the political power so constituted. Therefore, individual rights against the state "are not *originary* but rather emerge from a transformation of individual liberties that were at first *reciprocally* granted" (BFN 457)—that is, established in the very formation of their collective will.

I am not sure how neatly these two steps can be distinguished from one another, but I shall let that go. My concern lies with the first step. Observe that Habermas himself describes the horizontal sociation, which he holds to be foundational, as a process where citizens "recognizing one another as equals, mutually accord rights to one another" (BFN 457). Whence, I ask, comes their recognition of one another as equals? As Habermas's very words imply, this is not a commitment they acquire in virtue of such association, but rather one which defines the sort of association they understand themselves to be forming. This commitment amounts in fact to precisely what I have been calling the principle of respect. And that principle embodies an individual right—a right requiring that the terms of political association to be developed be as transparent to one's own reason as to that of others. Such a right does not arise through being mutually granted by citizens as they form their collective will. It has an independent authority which individuals must acknowledge if they are to form a democratic will in the way Habermas envisions.

I see no reason, therefore, to change my conclusion that Habermas's notion of radical democracy is not really so radical that it differs materially from the idea of liberal democracy to which he imagines himself opposed. Failing to note the moral basis of his own commitment to democratic self-rule, he slips past the fact that he too assumes, if only implicitly, the antecedent authority of individual rights.

## VIII. FREEDOM AND MORALITY

Modern democracy is government by self-rule, animated by the ideal that the basic rules of political life should draw their legitimacy from being the object of reasonable agreement among those whom they are to bind. But we misunderstand the nature of the democratic ideal, I have insisted, if we suppose that for it, the collective will of the citizens constitutes the ultimate source of authority. Looking more closely at the notion of the collective will employed, we see that it is conceived in moral terms. Popular sovereignty can be understood as manifesting itself through reasonable agreement, only if it is defined as heeding the obligation of respect for persons. Democracy is thus a moral conception, and not just in the trivial sense that the principles and values by which a democratic people organize their political life are recognizably moral in character. More profoundly, democracy involves commitment to a moral principle that citizens must see as binding on them independent of their democratic will. Respect for persons is what gives their democratic will the normative shape it has.

---

[8] I first presented this criticism in a review of *Faktizität und Geltung*, "Die Wurzeln radikaler Demokratie," *Deutsche Zeitschrift für Philosophie*, XLI (1993): 321-27. I developed it further in *Morals of Modernity*, chapter 10.

I have claimed that Habermas misses this moral basis of modern democracy and that Rawls, too, though recognizing it in part, fails to bring it out as distinctly as he should. My point, as I hope is plain, is not to reject the project of working out a conception of liberal democracy that is politically freestanding. That project I share. But I believe its moral presuppositions should be made explicit. Why, we may ask, should there be such reluctance to admit these assumptions? One reason is not far to seek. Liberal thinkers who continue to sympathize with earlier individualist versions of liberal democracy do not hesitate to expound the moral foundations of this ideal as they see them.[28] But both Habermas and Rawls, in their different ways, aim to work out a political conception that does not depend upon comprehensive moral views, especially individualist ones, about the ends of life. In the effort to devise a political conception that is thus freestanding, it is only too easy to slide over the moral commitments propelling this very enterprise.

I do think it important that the principle of respect does not express or entail a comprehensive moral philosophy. It has its place in a great many otherwise disparate ideas of the human good. Particularly significant is the fact that it can figure in those conceptions of life which refuse to accord supreme value to critical reflection and call instead upon forms of moral allegiance that are rooted in a sentiment of belonging. We may feel that our deepest commitments are constitutive rather than elective, fashioning our very sense of the persons we are and the choices we can understand ourselves as making, and at the same time believe that political principles, relying as they do upon coercion, must be reasonably acceptable to all whom they are to bind, even to those who differ with us on the relative importance of critical reflection. There exist, of course, conceptions of the human good which reject a principle of respect for persons. No doubt, many of them also desire to limit critical reflection and preach instead obedience to traditional, often religious authorities. But these two views do not go necessarily together. In fact, individualist views of life, too, when they exalt the will to power of creative individuals and heap contempt upon the bulk of mankind, can prove illiberal. A reverence for tradition is not inherently hostile to liberal ideals and more than the cultivation of individuality is intrinsically friendly to them.

---

[28] See, for example, Jeremy Waldron, *Liberal Rights* (New York: Cambridge, 1993), pp 56-57, 163-68.

But it also bears noting that reasonable people, exercising the basic capacities of reason and conversing with others in good faith, are not obliged on this basis alone to believe that the use of force should remain within the bounds of consensus. Being reasonable (in my sense of the term) does not entail the principle of respect for persons. If the moral basis of liberal democracy stands free of any comprehensive conception of the good, it does not on that account become inherently universal. Its appeal can extend only so far as people happen to be committed to the principle of respect.[27] We ourselves may find this commitment so obvious that, like Habermas and Rawls, we fail to recognize or mention it. We may simply look through it, as we reflect upon the constitutional rules a democratic people should give themselves to govern their political life. But in reality, this commitment forms the specific moral "we" we are, even as we remain divided by important differences about other fundamental matters. Self-understanding, here as elsewhere, requires that we turn our mind to what lies so close to us as to go unseen. Then we can comprehend why this moral outlook, transparent though it may be to us, is not universal and has met indeed with sincere rejection from others.

Yet I wonder whether there is not another reason why Habermas and Rawls do not make explicit the moral basis of their conceptions of liberal democracy. Perhaps in their case, as certainly in the case of others, another contemporary ideal has come to stand in the way of seeing our deepest commitments for what they are. In our time, freedom of self-determination, as both an individual and a collective value, enjoys a tremendous prestige. It commands so ready an allegiance that all other values can seem of subordinate importance. Individuals can be bound, it is said, only by the rules they give themselves. So, too, we continually hear that, in their collective capacity as members of a political community, citizens are to determine themselves the principles by which they will live.[28] I doubt that the moral life in general can make much sense if its roots are sought in freedom as a supreme value.[29] I am certain, as I have argued in this essay, that our commitment to democracy or political self-determination cannot be understood except by appeal to a higher moral authority, which is the obligation to respect one another as persons.

---

[27] I agree therefore with Stanley Fish—"Mission Impossible: Settling the Just Bounds Between Church and State," *Columbia Law Review*, XCVII, 8 (December 1997): 2255-333—that liberalism, like any political doctrine, must involve exclusion. All the same, a distinctive virtue of political liberalism is the breadth of its inclusiveness.

CHARLES LARMORE

University of Chicago

---

[28] Cf. Habermas, *Die Einbeziehung des Anderen*, p. 301: "Menschenrechte mögen moralisch noch so gut begründet werden können, sie dürfen aber einem Souverän [here, the democratic will, C.L.] nicht paternalistisch übergestülpt werden.... Dieser Idee widerspräche es, wenn der demokratische Verfassungsgesetzgeber die Menschenrechte als so etwas wie moralische Tatsachen schon vorfinden würde, um sie nur noch zu positivieren."

[29] For reflections along these lines, see *The Morals of Modernity*, pp 87-88, and "L'autonomie de la morale," *Philosophiques* (Québec), XXIV, 2 (automne 1997): 313-28.

# BOOK REVIEWS

*Mind, Language, and Society: Philosophy in the Real World.* JOHN SEARLE. New York: Basic Books, 1998. x + 173 p. Cloth $21.00.

John Searle describes this book as a summary explanation of his views on mind, language, and society, and on "how they fit into our overall contemporary conception of the universe." As such, it is a great success. Searle's prose is lucid, but vigorous and muscular. He has things to say, and he says them well. He also lays out a great deal of territory, so the book is eminently usable as an introduction not only to Searle's own ideas, but to many important themes in contemporary philosophy of mind and social philosophy. Anyone who has tried to write introductory books on these subjects will know how difficult it is to achieve that.

Searle starts with a chapter laying out the basic tenets of his metaphysics. He defends realism against idealism, skepticism, and relativism. His realism is composed of a number of "default" positions or things that are taken for granted, and forms what he calls the *background* of our thoughts and language. The background includes the presuppositions that there is a real world, that we can perceive it and know about it, that our words have clear meanings, that statements are true when they correspond with the way things are, and that causation is a real relation between events. Searle fends off idealist, relativist, or skeptical challenges to the background in fairly familiar ways. Indeed, he thinks that the arguments presented by those who advance such challenges are so feeble that they call for a different diagnosis, which he finds in the "will to power." If some kind of relativism or constructionism is true, then it is we who are in charge, rather than the world which is in charge of us.

Chapter 2 defends the existence of mind as a biological phenomenon, while chapter 3 continues with the structure of consciousness, including interesting material on the unity of consciousness. Subsequent chapters cover intentionality, the structure of the social universe, and the nature of language. On all these topics, Searle has interesting things to say, often convincingly, sometimes controversially.

The avowed centerpiece of the book is Searle's rejection of both materialism and dualism. It is this which underpins his claim to be giving a unified metaphysics. For once mind is firmly established in

0022-362X/99/9612/626–29

nature, the rest is, as it were, a question of mopping up. The intentional powers of the mind, and the construction of a complex social world that includes language, are placed on the secure foundation provided by a nonmysterious theory of consciousness. So everything depends on this. Yet in spite of the brisk confidence with which Searle proposes his theory, it is actually very hard to see just what it is, and harder still to see how it triumphantly solves the problems that bedevil others.

Searle begins by providing sympathetic accounts of the pressures that lead both to dualism and to materialism. In motivating dualism, he particularly stresses the inner, qualitative, and subjective nature of conscious states, or their "first-person ontology." It is this first-person ontology which is disregarded or distorted by "materialist," or objective, scientific, reductions of consciousness, such as behaviorism, functionalism, or physicalism. Since dualism and materialism conflict, the solution is to transcend the categories: "In short, the way to reply to materialism is to point out that it ignores the real existence of consciousness. The way to defeat dualism is simply to refuse to accept the system of categories that makes consciousness out as something nonbiological, not a part of the natural world" (52). The upshot is given in six propositions, which I quote in full:

> [1] Consciousness consists of inner, qualitative, subjective states and processes. It has therefore a first-person ontology. [2] Because it has a first-person ontology, consciousness cannot be reduced to third-person phenomena in the way that is typical of other natural phenomena, such as heat, liquidity, or solidity. [3] Consciousness is, above all, a biological phenomenon. Conscious processes are biological processes. [4] Conscious processes are caused by lower-level neuronal processes in the brain. [5] Consciousness consists of higher-level processes realized in the structure of the brain. [6] There is, as far as we know, no reason in principle why we could not build an artificial brain that also causes and realizes consciousness. [Format altered by author.]

Searle adds a closure clause: "that is it" (53).

But the matter cannot be left to rest there, since if it does the doubt remains that Searle has simply seized some attractive features of each of dualism and materialism, without doing anything to show how their combination is intelligible. Thus, consider proposition [5]. In explaining this, Searle uses a popular kind of example, claiming that consciousness is "a higher-level feature of the brain in the same way that digestion is a higher-level feature of the stomach" (52). Many philosophers of mind would applaud this comparison. But they would regard it as a vindication of some kind of physicalism or functionalism. Their mistake, says Searle, is to forget or ignore proposition [1]. But then the question is why, *if* [1] is incompatible with materialism, it is nevertheless compatible with [5]. Digestion, as Searle admits, is an objective, third-person phenomenon. All Searle says here is that it is just a fact that some higher-order states of the brain have a special subjective, qualitative, inner ontology. But he does nothing to tell us how that can be a fact.

The first-person ontological status of conscious states is clearly crucial here. On the one hand, it is beefy and strange enough to rule out functionalism and physicalism. On the other, it is not beefy enough to provide an insuperable mystery for the idea that conscious states are higher-order, causally efficacious, states of the brain. So it is a pity that Searle's introduction of these notions is quite perfunctory. Conscious states are "inner" just in the sense that they are states of the body and brain; they are "qualitative" in that it feels like something to be in such a state; and they are "subjective" in the sense that they are always experienced by a human or animal subject (42). Now, each of these can be taken as quite trivial consequences of the way we ascribe conscious states to persons, just parallel to the way in which dents are inner, qualitative, and subjective states of surfaces. It is not at all clear that these aspects of conscious states are beefy enough to refute functionalism and physicalism. And it is still less clear that they can do this without leaving Searle's own theory quite unintelligible, unless, as is so often the case, there are two different grades of beef in play. So I doubt if Searle's solution to the mind-body problem is really a solution at all. He may wish to avoid labels like 'dualism' and 'materialism', but this is not the same as avoiding the problems for which they are labels. We should be reminded here of the issue between John Locke, who thought that is was "God's good pleasure" to annex definite sensations to physical states, and Gottfried Leibniz, who complained that God did nothing arbitrarily, but that the mental must be a rationally intelligible expression of the physical. Searle writes as if it solved things just to remind ourselves that the physical does cause the mental. But while matters remain there, the crucial issue between Locke and Leibniz never swims into focus.

There are other places where Searle is a little kinder to his own ideas than to those of others. One concerns the place that "intrinsic intentionality" has in his system. According to Searle, some conscious states possess intrinsic directedness or intentionality. By contrast, other things, and notably sentences, possess intentionality that is only derived from the intrinsic intentionality of the conscious states of their speakers (93). Yet Searle also wants to accommodate the fact that to think many things seems to require mastery of a language. His solution is that "The child starts with simple prelinguistic intentionality. It then learns a simple vocabulary that enables it to have a richer intentionality, which in turn enables a richer vocabulary, which in turn enables a richer intentionality—and so on up in a boot-strapping process" (153). Again, how exactly is this supposed to work? Consider an example. Suppose, as Jonathan Bennett once argued, that nonlinguistic babies or animals cannot think thoughts whose content concerns a remote past. Now suppose that mastery of the linguistic device of tense rectifies that. How can it do so if the linguistic device itself possesses meaning that is entirely derived from the "intrinsic intentionality" of persons who use it? We seem to be faced not so much with a bootstrap as with a vicious circle: the meaning of the devices of tense both creates and depends upon the content of thought of the people who use it. There may be a way out of the circle, but it is not at all clear what it would be, and Searle is content to leave it there.

While such problems beset Searle's system, they should not be allowed to mask his real achievement in bringing some very difficult philosophy forcefully and vividly into the popular arena.

SIMON BLACKBURN

University of North Carolina/Chapel Hill

*Consequences of Enlightenment.* ANTHONY J. CASCARDI. New York: Cambridge University Press, 1999. viii + 268 p. Cloth $54.95, paper $19.95.

Anthony Cascardi's *Consequences of Enlightenment* has three principal themes. First, as his title suggests, he is concerned about the legacy of the Enlightenment, specifically in connection with the Frankfurt-school tradition of critical theory. That tradition is thought to have encountered an impasse: its critique of instrumental reason seems so thorough that it leaves no room for noninstrumental rationality, yet such rationality is needed for the critique of instrumental reason to be possible in the first place. The impasse is, of course, not unique to critical theory, as it mars any critique of reason that aims to be systematic—hence the relevance of Immanuel Kant throughout Cascardi's book. His strategy is to analyze ways in which political critique can be "reevaluated in light of an aesthetic critique modeled along lines sketched out in Kant's *Critique of Judgment*" such that this impasse can be avoided (1-2). Second, he discloses the Kantian roots of a number of modern and postmodern thinkers: Jürgen Habermas, Slavoj Zizec, Hannah Arendt, Jean-François Lyotard, Chantal Mouffe, Roberto Unger, Theodor Adorno, Richard Rorty, Roland Barthes (just to give a partial list of the theorists Cascardi discusses, only some of whom are Kantians). It turns out that Kant's notion of aesthetic judgment, characterized by indeterminacy, is central in their respective conceptions of political critique. Third, as Cascardi traces these theorists to the *Critique of Judgment* while focusing on their political philosophies, he discloses a deep link in contemporary continental philosophy between aesthetics and politics based on a broad conception of aesthetics, that is, one which is not confined to autonomous art but which, following a reading of Kant, encompasses the subjective realm of affectivity that is tied to a universal common sense. Aesthetic critique is thus the single issue running through all three themes. (In fact, *Aesthetics as Critique* was originally the subtitle of Cascardi's book, but it was left off, he said, in the final stage of publication.)

Yet what does aesthetics—particularly the indeterminacy of aesthetic judgment–contribute to our understanding of contemporary political critique? Addressing this question is a direct way of ascertaining whether Cascardi provides the conceptual resources to advance the debate about critique he has opened up. While the concept of indeterminacy indeed has aesthetic and political consequences, I am afraid they are not always the ones Cascardi envisions.

I

Cascardi's aim is to explore the indeterminacy that attaches to aesthetic judgment, starting from the following problem. There is no objective principle of taste; rather, there is only a "subjective principle—...the indeterminate idea of the supersensible within us" (68). Such a subjective, indeterminate principle should not be thought to render aesthetic judgment deficient, as it often has been, but should rather be seen as the basis of a unique form of judgment capable of opening up affective riches which have political consequences for the exercise of critique. Cascardi thus begins with an epistemological problem about aesthetic judgment, yet ends up with an argument about the political consequences of the "solution" to that problem. But how does a *critique of aesthetics* in the Kantian tradition lead to a notion of *aesthetics as critique* in postmodernity? Cascardi's path from the indeterminacy of aesthetic judgment to politics takes us through the "larger project of boundary distinction, which is a concern that governs the entire Kantian critical project" (73). He is referring here to the separation between cognition (*Critique of Pure Reason*) and morality (*Critique of Practical Reason*), and to the fact that the *Critique of Judgment* is supposed to provide the mediating concept, judgment, between theoretical and practical reason. What relates the issue of the relations among Kant's three Critiques to the issue of the indeterminacy of aesthetic judgment (and what thereby establishes the broader notion of aesthetics) is that the division between cognition and morality (and, even more so, the desire for their reintegration) is not something we *know* but is rather something that is "affectively apprehended" by the subject in the form of aesthetic judgment (44). To grasp aesthetic judgment in its indeterminacy is to grasp this realm of affectivity as a source of critique (as well as of possible mediation).

But while it is clear that indeterminacy is a central notion for Cascardi, there seems to be some slippage or play among the following senses of it which, in turn, are tied to an equally important, though undeveloped concept of contingency which also has multiple meanings. Indeterminacy may be true to its name here, but at the same time, it may reveal the contingency of Cascardi's own critical project.

(1) The *indeterminacy of aesthetic judgment*. As we saw, aesthetic judgment is indeterminate because its principle is the "indeterminate idea of the supersensible within us." In Kant's language, it is re-

flective (indeterminate) rather than determinate because it has no determinate concepts, which means it cannot convey any knowledge (61-62). It is also subjective, as it is tied to a feeling of pleasure, yet it is still thought to be universal. With such an account of aesthetic judgment, it is unclear—or indeterminate (perhaps more for Cascardi than for Kant)—how it works, whether it is indeed universal, or what its philosophical role is.

(2) The *indeterminacy of art.* Cascardi insists, following Kant, that the nature and identity of works of art be left unspecified (55); all we can say is that they are created by genius. So it is indeterminate how something becomes or is a work of art.

(3) The *indeterminacy of the aesthetic domain.* Because aesthetic judgment and art are indeterminate, the aesthetic domain of philosophy is without rules, principles, or criteria (16, 59). So it is indeterminate what does or does not belong within this domain, especially when it is extended (by Cascardi) beyond the realm of autonomous art.

(4) The *indeterminacy of cognition and morality.* The Kantian system has a "constitutive opacity" (91) because (as we saw) reason cannot account for the division or unity of cognition and morality. Such an account lies beyond the limits of reason, yet it is crucial for its legitimate exercise (this is the Kantian counterpart of critical theory's impasse). What makes reason *in*determinate once its limits become evident is that we do not *know* these limits through the use of determinate concepts; rather, we ascertain them without concepts by feeling what lies beyond them. These feelings are reflected in aesthetic judgment, the structure of which is as indeterminate as the feelings it mirrors.

To complicate matters further, these four senses of indeterminacy are linked to two different senses of *contingency.* (A) The *contingency of the cognition/morality distinction.* Although aesthetic reflective judgment is supposed to be the middle term linking cognition and morality, it necessarily fails in this capacity precisely because it is indeterminate (72-77). Such failure shows that the divisions Kant establishes between cognition and morality are contingent (57), as is any unity of reason he might claim. Cascardi argues, however, that something very positive follows from this failure and contingency; for the appropriate philosophical response here is to recognize that there is no "'solution' to the division of cognition and morality or a proof of the unity of experience" (89). What we have instead is "an antinomy: that the principle of the universal validity of [judgments] of taste must indeed exist, but that the principle cannot be *shown* to exist" (89). Thus, while judgment fails as a mediator, its failure is its strength; for in trying to serve as a mediator, judgment establishes that it is legitimate even though it cannot be proven by reason, because we need judgment to complete the critique of reason, without which completion we cannot exercise reason legitimately. Thus the legitimacy of judgment is intimately tied up with that of reason itself.

(B) The *contingency of the world.* The "fact" that the world is not structured according to any rational purposes is embodied in human affects, specifically pleasure/pain, and is subsequently reflected in aesthetic judgments. In short, our affects and judgments are traces of the world's contingency. But what type of contingency are we dealing with here? That of history or of nature? Cascardi is not clear on this key issue, but it is likely he means both, given what he sees (below) as stemming from contingency. In fact, he says little else about contingency, which is rather surprising since it clearly plays an important role in his argument; for Cascardi claims that what Kant seeks in arguing for the universality of subjectively grounded aesthetic judgment is "a reappraisal of the role of the passions in revealing how human purposes fit in relation to the contingency of the world" (55).

## II

Now, to mimic the title of Cascardi's book, what are the enlightenment consequences of indeterminacy and contingency? First, indeterminacy serves as the model for Cascardi's notion of subjectivity: "The indeterminacy of the domain that Kant couches in affect thus becomes the basis for a conception of the subject as a site of loss or dispossession rather than for the recuperation of a positive relationship to the world" (11). Second, contingency underlies his notion of affectivity, for he claims that affects, such as pleasure or pain, are "in the realm of the subject, a trace of the contingency of the world" (48, 10). Combining these two points, the relationship between indeterminacy and contingency seems to be something like the following: the *contingency* of the world is experienced (even suffered) by the subject, and such affective experience is mirrored in *indeterminate* aesthetic reflective judgments. Third, Cascardi says that critical force *attaches* to the indeterminacy of the aesthetic (174, 57); that is, it is possible "to locate a critical value in a form of judgment that takes its bearings by the affects and whose procedures are fundamentally reflective" (55). Fourth, one, if not *the,* major consequence of Cascardi's multiple forms of indeterminacy and contingency is that they are crucial to political transformation (179), for they open up (public) space for politics (and ethics). The link between the aesthetic and the political is provided by the common sense or affectivity that grounds the universality of indeterminate judgment, for "the affects provide an opening for an ethical determination of the subject—ethical [and later also political] because determined by a responsibility to the other" (11). Affectivity is what is universally communicable across the aesthetic, ethical, and political spheres of indeterminate judgment.

The aesthetics-politics link is most concrete in Cascardi's characterization of democracy, which brings together his notions of subjectivity, affectivity, and critique. He argues that contemporary thinkers "interested in radicalizing democracy," himself included, "have turned to the project of an aesthetically grounded critique as part of an effort to identify what the notions of communication and consensus seem to leave out: the ongoing desire of subjects to refashion the structures they inhabit, coupled with the desire to perpetuate democracy by recognizing that Enlightenment democratic universalism and its institutions were structurally, inherently incomplete" (179). The resulting notion of democracy is as follows: "Democracy describes that form of politics suited to a subject that recognizes itself as having no essential nature, that refuses to understand itself as the effect of some external or underlying cause, and that regards change as inevitable insofar as closure is unattainable" (180). That is, just as we lack an essential or fixed nature, our political principles and institutions are constitutively incomplete (174, 194); as a result, change and revision are the norms in politics. Political critique is public reflection on what changes and revisions are desirable and possible. Modeled on the notion of aesthetic judgment, such critique should not focus on achieving consensus on determinate conceptions of liberal principles, such as "freedom" or "equality" (180). Rather, political critique should be guided by the (Lyotardian) question: "What becomes of the polis when the power of judgment is trained not upon some determinate object, but rather on what the determination of any object leaves out"—from, for example, feminist, racial, or ecological perspectives (181, 174; cf. 201-04)? Identification of such exclusion is the new political focus instead of the ideal of consensus which, under conditions of contingency, can never be attained. Political judgment thus mirrors aesthetic judgment in that both are indeterminate and critical. Moreover, Cascardi claims they are critical *because* they are indeterminate. We have thus arrived at the goal of his argument: the indeterminate structure of aesthetic judgment is said to establish a subjective, affective basis for political critique aimed at creating a democratic analog of the universal common sense that is both presupposed, and created, by aesthetic judgments.

### III

Adorno once remarked that while a work of art appears to be determinate when it is political, it is in fact indeterminate, though not for the reasons Cascardi provides. Rather, the work that appears to be determinate is actually indeterminate because it has been determined by so many historical factors that, with respect to their causality, are hard to pin down. Carrying Adorno's dialectic one step further, the work appears to be indeterminate because of these many factors, yet they are what make it determinate. So the (political) work of art that first appears to be determinate turns out to be indeterminate, yet it appears to be indeterminate only because it is determinate. Analyzing these appearances instead of defending either determinacy or indeterminacy alone is the difficult task shared by aesthetic and political critique. In this light, if art is to be understood in terms of the notion of indeterminacy at all, then I would suggest that it be understood in terms of the notion of *indeterminate determinacy*. By this notion, I mean a determinacy which is shaped by issues definitive of the history of the present but which is open to multiple interpretations so that it escapes the stranglehold of reigning ideologies and is thus able to engage them; critique lies in the determinate space between this escape and engagement. Such indeterminate determinacy is to be distinguished from the notion of *determinate indeterminacy*, which is a commitment in art (often later mirrored in aesthetics) to be indeterminately related to, and thus distant from, history, as if art's ability to be critical would lie in such indeterminacy or distance alone. This strikes me as the kind of indeterminacy Cascardi embraces in his accounts of aesthetic as well as political judgment. But I think such indeterminacy is a symptom of art's historical irrelevance couched in the reputedly redeeming concept of indeterminacy, as if an appeal to that concept could transform irrelevance (a fact) into a virtue (a norm). Art does not need to be redeemed in its irrelevance. Rather, like aesthetic and political judgment, it has to find a path to relevance that grounds critique. With this point in mind, I would like to raise several questions about Cascardi's treatment of indeterminacy in connection with aesthetics and politics.

(1) The ground of Cascardi's aesthetics-politics link is the duo of (epistemological) indeterminacy and (ontological) contingency. But how could they possibly anchor political critique? Critique seems incapable of supporting any concrete political discourse, so long as it is grounded in nothing other than indeterminacy, even when indeterminacy is rendered more concrete by being tied to the notion of exclusion. But my criticism of Cascardi's position is *not*

that political critique is not grounded philosophically; for I agree that it is groundless in the sense that it is contingent (240) and that, at the same time, it is grounded as much as it needs to be. My criticism is rather that what he provides through his account of contemporary political critique is a description (albeit an accurate one) of its predicament, namely, that it is indeterminate. The reason it is indeterminate, however, is not that it is modeled on aesthetic judgment, but because it lacks a determinate foothold in politics. The problem here, therefore, is not a philosophical one; if it were, the concept of indeterminacy may very well be a solution, since indeterminate judgment is (self-) grounded. The problem of political critique Cascardi describes is rather that it lacks a *political* grounding, a problem to which no philosophical concept alone could ever be the solution. The link here with contemporary aesthetics is that, with few exceptions, it is in a similar predicament: it lacks a determinate foothold in art. In this light, indeterminacy is the name of a problem shared by aesthetic and political judgment rather than a resource for solving it. To make this point using the earlier notion of exclusion, what is excluded here is either art or politics.

(2) Cascardi inscribes into the ontology of art a certain impossibility which is expressed by Jacques Lacan, albeit in Cascardi's words, as follows: "the category of 'art' comes into being within the culture of enlightened modernity not as the effect of any determinate set of causes but only as the consequences of an impossibility—specifically, as a consequence of constituting the 'whole' of experience by bridging the 'two worlds' that Kant had set aside in his first two critiques" (98). The impossibility of ever unifying experience, which is a failure of reason, is thus the origin of art (and of judgments thereof, since such failure has the effect, as we have seen, of validating aesthetic judgment). Yet what is the rationale for inscribing such impossibility, which, after all, is an epistemological matter, into the ontological heart of art? Why should art be ontologically determined to be indeterminate because of an epistemological limitation of reason? Are the determinate causes of art really so obscure, however complex they may be? Instead of addressing such questions, Cascardi has written his account of aesthetic judgment into his account of art. The result is a conception of art defined by aesthetics conceived as "independent of the specificity of works of fine art" (2-3). Why should art be defined by an aesthetics deliberately separated from art? If Cascardi is going to appeal to aesthetics to understand politics, I think he might first rethink art and then reconnect aesthetics to art.

(3) Of course, Cascardi makes it clear that he is disinterested in art because aesthetic judgment has relevance and application well beyond art. That may be true, but he might first reconsider his disinterest because it has negative consequences for his own strategy. He explicitly works not only with a notion of aesthetics which is not confined to autonomous art but which is also separated from it. This means that aesthetic judgment (and critique) is likewise separated from art. Then what is aesthetic judgment a judgment *of*? Affectivity? Besides being rather indeterminate, that answer would not work because (the universal communicability of) affectivity is the ground of aesthetic judgment and so cannot be its object. Then what is? Contingency? It is not clear. This unclarity is a problem for Cascardi because he models political judgment on aesthetic judgment and is thus faced with a parallel problem: political judgment is analogously (if inadvertently) conceived as "independent of the specificity" of politics and is thus as separated from politics as aesthetic judgment is from art. The consequence of such separation is irrelevance, and indeterminacy is its symptom. So indeterminacy is not, as Cascardi claims, what opens up the possibility of (aesthetic or political) critique. I am not suggesting that it (or he) has necessarily closed off this possibility but that by elevating indeterminacy to both a philosophical principle and a political norm, he has mistaken a symptom of critique's irrelevance for a condition of its possibility. In a word, he has misidentified a symptom (indeterminacy) of the fact that we lack something (critique) as the norm to which we should aspire, hoping it (indeterminacy) will eliminate the lack. That would be like identifying the hunger we feel when we lack food as the norm to which we should aspire, hoping it will alleviate our hunger despite the continued absence of food.

MICHAEL KELLY

Columbia University

*In the Light of Logic.* SOLOMON FEFERMAN. New York: Oxford University Press, 1998. xii + 340 p. Cloth $49.95.

In December of 1998, a conference was held at Stanford University in honor of Solomon Feferman's seventieth birthday. Mathematicians, philosophers, computer scientists, historians, and even a linguist or two gathered from about a dozen countries to celebrate Feferman's work in logic. Surveying this work, one detects a reassuring uniformity of method, and a general template for research in the field: first, find an issue relevant to the philosophy or history of mathematics that calls for clarification; next, design formal representations of the relevant mathematical phenomena, and apply the methods of logic and proof theory to their analysis; finally, interpret the results of the inquiry with respect to the original issues. Feferman's contributions to mathematical logic are substantial, but his work is perhaps best appreciated in this broader context. It is precisely this interplay between mathematical logic and the philosophy of mathematics that constitutes the central theme of *In the Light of Logic*, a collection of reflective essays spanning the last two decades.

These writings do not offer a grand unified theory of mathematics, but, rather, focus on more sharply delimited issues. For example, a number of essays, including "Infinity in Mathematics," "Weyl Vindicated," and "Why a Little Bit Goes a Long Way," address the notion of *predicativity*. In an informal sense, the word is used to describe approaches to mathematics in which concepts are introduced "from the bottom up," in accordance with the Poincaré-Russell Vicious Circle Principle. From such a standpoint, one may be willing to accept quantification over the natural numbers as unproblematic; but, taking sets of natural numbers to be given by explicitly defined properties, one may reject definitions that treat the collection of all such sets as a completed totality. This informal stance forms the basis of Hermann Weyl's landmark 1918 work, *Das Kontinuum*,[1] in which a significant portion of real analysis is developed on such a foundation. Since then, a good deal of effort has gone toward providing a formal analysis of predicativity, but no one has done more than Feferman to shape our current understanding of the term. In the 1960s, he identified a constellation of formal theories that seem to

---

[1] Leipzig: von Veit, 1918.

0022-362X/99/9612/638–42

exhaust the predicativist's means, and using the methods of Gerhard Gentzen, Kurt Schütte, and the "Munich School" of proof theory, he determined the ordinal, $\Gamma_0$, that represents the least upper bound to their proof-theoretic strength. In more recent years, a number of theories have been shown to be proof-theoretically reducible to Feferman's, and most interested parties are willing to use this criterion to obtain a working definition of predicativity. In other words, the general feeling is that Feferman has got it right, both in providing a formal analysis of a certain foundational stance, and in identifying an interesting and significant segment of mathematical reasoning.

Ongoing work in proof theory has shown that a good deal of mathematics can be carried out in predicative theories. In fact, many mathematical arguments can be formalized in theories that are weaker; and sometimes one can even get away with theories conservative over primitive recursive arithmetic, which many take to represent David Hilbert's notion of *finitistic* reasoning. One route to obtaining results like these—dating back to Hilbert and Paul Bernays and vigorously pursued by Harvey Friedman, Stephen Simpson, and others—is to use restricted subsystems of second-order arithmetic to carry out the formalization. A drawback of this approach is that in these theories, mathematical objects (functions, fields, metric spaces, measures, and so on) have to be coded as sets of numbers, which explicitly restricts the discourse to mathematical objects that are countably presented. Feferman has developed a more flexible approach with his theories of explicit mathematics, where the countable interpretations lie, in a sense, behind the scenes. This difference aside, however, the two approaches paint pictures that are essentially in agreement as to "where" everyday mathematical reasoning takes place.

Some of the conclusions that Feferman wishes to draw from this analysis are controversial. While Friedman and others have discovered examples of combinatorial theorems (like Kruskal's theorem, and finitizations thereof) whose proofs are necessarily impredicative, Feferman has conjectured, on the basis of the evidence just described, that at least those branches of mathematics which have direct scientific applications can be developed on a predicative basis. He takes this to weigh against arguments offered by W. V. Quine and Hilary Putnam in favor of a realist, set-theoretic ontology, which maintain that such a mathematical ontology is indispensable to the sciences. Feferman laments the fact that arguments like these have, to a large extent, supplanted alternative philosophies for the foundations of mathematics, like constructivism or predicativism. In an essay provocatively titled "Infinity in Mathematics: Is Cantor Necessary?" he quotes Weyl in characterizing the Platonism which underlies Cantorian set theory as the "medieval metaphysics of mathematics," and pronounces it "utterly unsatisfactory as a philosophy of our subject" (248).

Feferman does not offer much detail as to the form of Platonism he takes to underlie set theory, however, nor does he discuss a number of philosophical views, realist or otherwise, that can reasonably be associated with the subject. His efforts are directed, more specifically, toward a critique of the indispensability arguments mentioned above. Suggesting that the Cantorian world view is dispensable is a sure-fire way of raising set theorists' hackles, and so it is important to keep clear what is at stake. Feferman is not primarily concerned with offering a normative account of mathematical practice, but, rather, determining the appropriate logical and philosophical analysis thereof. A set theorist could respond to the proof-theoretic account in a number of ways. First, she could argue that this account has something of a revisionist tinge: mathematicians use "large" sets and impredicative definitions freely, and though (after the fact, and in principle) we can often eliminate these features from a given theory, the impredicative conception may have been essential to the theory's development, and may still be essential to a full understanding of its terms. As far as the outer reaches of set theory are concerned, she could defend the subject as a search for principles that might, in the future, play a central role in everyday mathematics. After all, large cardinal axioms have consequences for the continuum, and Friedman has long been searching for interesting finitary consequences as well. If our scientific theories do not currently warrant granting large cardinals full ontological status, one day they might; and a liberal and open-minded view of mathematics should prevent us from adopting a philosophical stance that precludes this possibility. Along these lines, our set theorist might even question the extent to which a minimal mathematical ontology is desirable. If, as Georg Cantor and Hilbert maintained, the essence of mathematics lies in its freedom of abstraction, then the proof theorist's drive to tether the subject to a restricted ontology misses the point; the resulting account of mathematics may soothe the apprehensive philosopher, but it may have little to do with mathematics itself.

Feferman would, no doubt, have much to say in reply. But I think it is important to emphasize that there is more to the proof-theoretic program than the negative task of reining in profligate ontologies. A pair of essays titled "Foundational Ways" and "Working Foundations" offer a thoughtful and spirited discussion of research in the foundations of mathematics, surveying a range of methodological approaches, emphasizing the positive gains of such research, and illustrating these gains with well-chosen examples. The benefits discussed include the clarification of fundamental concepts, the resolution of conceptual problems and paradoxes, the systematic organization of mathematical ideas, and the avenues toward generalization made possible by such a presentation. From this point of view, the proof-theoretic formalizations and reductions discussed above are to be valued for the additional information they provide with regard to the structure of the mathematical theories under consideration. For example, the development of real analysis in a restricted framework helps us determine which features of the mathematical universe are essential to the subject, and how they are used; it shows us that the theorems of analysis hold true not only in the set theorist's well-stocked universe, but in restricted models as well; and it helps us locate the "concrete" content of the theory's abstract concepts. As to the broader philosophical implications, Feferman writes in "What Rests on What?":

> It seems to me that the information provided by the kind of case studies described here involving the formalization of considerable tracts of everyday mathematics in appropriate systems, in combination with the results of their proof-theoretic reductions...must be taken into account in the continuing effort to develop a relevant and sustainable philosophy of mathematics (207).

But with characteristic caution, he adds:

> Though I personally believe that the kinds of results described here on the whole strengthen the case for a nonplatonistic philosophy of mathematics and further undermine the case for set-theoretical realism, they do not speak for themselves to that extent, and it is at that point that a well-informed philosophical discussion must take over *(ibid.)*.

In other words, the results of the proof-theoretic inquiry seem to tell us something about mathematics, and should be taken into account; but the philosophical mileage one obtains from these results may vary.

*In the Light of Logic* is divided into five parts. The first part focuses on historical topics and surveys some of the major lines of thought that have shaped the development of mathematical logic in this century. The second part provides a general discussion of the foundational enterprise and includes a balanced critique of Imre Lakatos's views on mathematics; with regard to the latter, Feferman finds "much to agree with in both the general approach and detailed analysis," but ultimately concludes that the theory taken in isolation fails to account for the distinctive character of mathematics (78). The third part examines the work of Kurt Gödel and weaves together aspects of his life, mathematics, and philosophy in a moving way. The fourth part outlines the proof-theoretic view of mathematics embodied in the kinds of results discussed above; and the fifth explores the consequences of these results for the philosophy of mathematics, largely vis-à-vis indispensability arguments. With prose that is crisp and precise yet accessible to a general audience, this book is essential reading for anyone interested in these topics,.

For the most part, the research discussed in these essays is foundational in the traditional sense: the logician's task is seen to be that of devising formal accounts that characterize the basic stuff of which the mathematical universe is composed, and the philosopher's task is seen to be that of providing epistemological and ontological justifications of mathematics, as portrayed by these accounts. Today, there is a growing sense that these are not the only issues in the philosophy of mathematics which are worthy of sustained inquiry. For example, one may seek accounts of mathematics which illuminate the structural view of the subject that has been on the rise since the Cantor-Dedekind revolution of the last century; theories of meaning and rationality that explain how mathematical theories grow and evolve over time; or theories of discovery and explanation in mathematics, akin to those which have been developed for the sciences. But the tone of Feferman's writing suggests that he would not be averse to investigations of this sort, and traditionalists among us like to think that the general methodological approach exhibited in these essays has much to offer in that regard. In the end, the extent to which mathematical approaches can inform philosophical inquiry is subject to debate; but the rigor and clarity of Feferman's essays make a strong case on their behalf.

JEREMY AVIGAD

Carnegie Mellon University

# INDEX TO VOLUME XCVI

## INDEX OF SUBJECTS DISCUSSED IN THE JOURNAL

**Analytic/Synthetic**

Sher, Gila: Is There a Place for Philosophy in Quine's Theory? 96 (10): 491-524 '99

**Artifacts**

Millikan, Ruth Garrett: Wings, Spoons, Pills, and Quills: A Pluralist Theory of Function. 96 (4): 191-206 '99

**Bayesian Theory**

Christensen, David· Measuring Confirmation. 96 (9): 437-61 '99

Lange, Marc: Calibration and the Epistemological Role of Bayesian Conditionalization. 96 (6): 294-324 '99

**Belief**

Christensen, David: Measuring Confirmation. 96 (9): 437-61 '99

Hetherington, Stephen: Knowing Failably. 96 (11)· 565-87 '99

**Biology, Philosophy of**

Millikan, Ruth Garrett: Wings, Spoons, Pills, and Quills: A Pluralist Theory of Function. 96 (4): 191-206 '99

**Bratman, Michael**

Holton, Richard: Intention and Weakness of Will. 96 (5): 241-62 '99

**Carnap, Rudolf**

Sher, Gila: Is There a Place for Philosophy in Quine's Theory? 96 (10). 491-524 '99

**Chomsky, Noam**

Hintikka, Jaakko: The Emperor's New Intuitions. 96 (3). 127-47 '99

**Churchland, Paul M.**

Fodor, Jerry: All at Sea in Semantic Space: Churchland on Meaning Similarity. By Jerry Fodor and Ernie Lepore. 96 (8): 381-403 '99

Lepore, Ernie: All at Sea in Semantic Space Churchland on Meaning Similarity. By Jerry Fodor and Ernie Lepore. 96 (8): 381-403 '99

Tiffany, Evan: Semantics San Diego Style. 96 (8): 416-29 '99

**Cognitive Science**

Fodor, Jerry: All at Sea in Semantic Space: Churchland on Meaning Similarity. By Jerry Fodor and Ernie Lepore. 96 (8): 381-403 '99

Lepore, Ernie: All at Sea in Semantic Space: Churchland on Meaning Similarity. By Jerry Fodor and Ernie Lepore. 96 (8): 381-403 '99

Tiffany, Evan: Semantics San Diego Style. 96 (8): 416-29 '99

**Conspiracy**

Keeley, Brian L.: Of Conspiracy Theories. 96 (3): 109-26 '99

**Davidson, Donald**

Williams, Michael: Meaning and Deflationary Truth 96 (11)· 545-64 '99

**Determinism**

Kane, Robert: Responsibility, Luck, and Chance Reflections on Free Will and Indeterminism. 96 (5): 217-40 '99

**Epistemology**

Christensen, David: Measuring Confirmation. **96** (9): 437-61 '99
Douven, Igor: Putnam's Model-theoretic Argument Reconstructed. **96** (9): 479-90 '99
Goldman, Alvin I.: Internalism Exposed. **96** (6): 271-93 '99
Hetherington, Stephen: Knowing Failably. **96** (11): 565-87 '99
Jones, Karen: Second-hand Moral Knowledge. **96** (2): 55-78 '99
Keeley, Brian L.: Of Conspiracy Theories. **96** (3): 109-26 '99
Sher, Gila: Is There a Place for Philosophy in Quine's Theory? **96** (10): 491-524 '99
Sorensen, Roy: Seeing Intersecting Eclipses. **96** (1): 25-49 '99

**Ethics**

Hershenov, David B.: Restitution and Revenge. **96** (2): 79-94 '99
Holton, Richard: Intention and Weakness of Will. **96** (5): 241-62 '99
Jones, Karen: Second-hand Moral Knowledge. **96** (2): 55-78 '99
Kane, Robert: Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism. **96** (5): 217-40 '99

**Evaluation**

Morreau, Michael: Supervaluation Can Leave Truth-value Gaps after All. **96** (3): 148-56 '99

**Explanation**

Keeley, Brian L.: Of Conspiracy Theories. **96** (3): 109-26 '99

**Field, Hartry**

Williams, Michael: Meaning and Deflationary Truth. **96** (11): 545-64 '99

**Fodor, Jerry**

Morreau, Michael: Supervaluation Can Leave Truth-value Gaps after All. **96** (3): 148-56 '99

**Free Will**

Kane, Robert: Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism. **96** (5): 217-40 '99

**Functions**

Millikan, Ruth Garrett: Wings, Spoons, Pills, and Quills: A Pluralist Theory of Function. **96** (4): 191-206 '99

**Game Theory**

Alexander, Jason: Bargaining with Neighbors: Is Justice Contagious? By Jason Alexander and Brian Skyrms. **96** (11): 588-98 '99
Skyrms, Brian: Bargaining with Neighbors: Is Justice Contagious? By Jason Alexander and Brian Skyrms. **96** (11): 588-98 '99

**Gettier, Edmund L.**

Hetherington, Stephen: Knowing Failably. **96** (11): 565-87 '99

**Gödel, Kurt**

Azzouni, Jody: Comments on Shapiro. **96** (10): 541-44 '99
Field, Hartry: Deflating the Conservativeness Argument. **96** (10): 533-40 '99
McCall, Storrs: Can a Turing Machine Know that the Gödel Sentence Is True? **96** (10): 525-32 '99

**Haecceitism**

Huggett, Nick: Atomic Metaphysics. **96** (1): 5-24 '99

**Habermas, Jürgen**

Larmore, Charles: The Moral Basis of Political Liberalism. **96** (12): 599-625 '99

**Holism**

Sher, Gila: Is There a Place for Philosophy in Quine's Theory? **96** (10): 491-524 '99

**Horwich, Paul**

Williams, Michael: Meaning and Deflationary Truth. 96 (11): 545-64 '99

**Infinity**

Almog, Joseph: Nothing, Something, Infinity. 96 (9): 462-78 '99

**Internalism**

Goldman, Alvin I.: Internalism Exposed. 96 (6): 271-93 '99

**Intuition**

Hintikka, Jaakko: The Emperor's New Intuitions. 96 (3): 127-47 '99

**Justice**

Alexander, Jason: Bargaining with Neighbors: Is Justice Contagious? By Jason Alexander and Brian Skyrms 96 (11). 588-98 '99
Hershenov, David B.: Restitution and Revenge. 96 (2). 79-94 '99
Larmore, Charles: The Moral Basis of Political Liberalism. 96 (12): 599-625 '99
Skyrms, Brian: Bargaining with Neighbors: Is Justice Contagious? By Jason Alexander and Brian Skyrms. 96 (11). 588-98 '99

**Kripke, Saul**

Hintikka, Jaakko: The Emperor's New Intuitions 96 (3): 127-47 '99

**Language, Philosophy of**

Fodor, Jerry: All at Sea in Semantic Space: Churchland on Meaning Similarity. By Jerry Fodor and Ernie Lepore. 96 (8): 381-403 '99
Lepore, Ernie: All at Sea in Semantic Space: Churchland on Meaning Similarity. By Jerry Fodor and Ernie Lepore. 96 (8): 381-403 '99
Morreau, Michael: Supervaluation Can Leave Truth-value Gaps after All. 96 (3): 148-56 '99
Sher, Gila: Is There a Place for Philosophy in Quine's Theory? 96 (10): 491-524 '99
Sider, Theodore: Presentism and Ontological Commitment. 96 (7): 325-47 '99
Williams, Michael: Meaning and Deflationary Truth. 96 (11): 545-64 '99

**Lepore, Ernie**

Morreau, Michael: Supervaluation Can Leave Truth-value Gaps after All. 96 (3): 148-56 '99

**Levi, Isaac**

Lange, Marc: Calibration and the Epistemological Role of Bayesian Conditionalization. 96 (6): 294-324 '99

**Liberalism**

Larmore, Charles: The Moral Basis of Political Liberalism. 96 (12): 599-625 '99

**Logic**

Azzouni, Jody: Comments on Shapiro. 96 (10): 541-44 '99
Field, Hartry: Deflating the Conservativeness Argument. 96 (10): 533-40 '99
Katz, Bernard D.: On a Supposed Counterexample to Modus Ponens. 96 (8): 404-15 '99

**Logic, Philosophy of**

Azzouni, Jody: Comments on Shapiro. 96 (10): 541-44 '99
Field, Hartry: Deflating the Conservativeness Argument. 96 (10): 533-40 '99
Yi, Byeong-uk: Is Two a Property? 96 (4): 163-90 '99

**Luck**

Kane, Robert: Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism. 96 (5): 217-40 '99

**Mathematics, Philosophy of**

Yi, Byeong-uk: Is Two a Property? 96 (4): 163-90 '99

**Meaning**

Williams, Michael: Meaning and Deflationary Truth. 96 (11): 545-64 '99

**Meaning, Theory of**

Morreau, Michael: Supervaluation Can Leave Truth-value Gaps after All. 96 (3): 148-56 '99

**Metaphysics**

Almog, Joseph: Nothing, Something, Infinity. 96 (9): 462-78 '99
Huggett, Nick: Atomic Metaphysics. 96 (1): 5-24 '99
Sider, Theodore: Presentism and Ontological Commitment. 96 (7): 325-47 '99
Yi, Byeong-uk: Is Two a Property? 96 (4): 163-90 '99

**Mind, Philosophy of**

Millikan, Ruth Garrett: Wings, Spoons, Pills, and Quills. A Pluralist Theory of Function. 96 (4): 191-206 '99

**New Books**

96 (4): 211-16
96 (5): 270
96 (7): 374-80
96 (8): 436

**New Books: Anthologies**

96 (3): 157-62

**Notes and News**

96 (3): 162

**Nothing**

Almog, Joseph: Nothing, Something, Infinity 96 (9): 462-78 '99

**Ontology**

Sider, Theodore: Presentism and Ontological Commitment. 96 (7): 325-47 '99

**Parmenides**

Almog, Joseph: Nothing, Something, Infinity. 96 (9): 462-78 '99

**Perception**

Sorensen, Roy: Seeing Intersecting Eclipses. 96 (1): 25-49 '99

**Philosophy, Analytic**

Hintikka, Jaakko. The Emperor's New Intuitions. 96 (3): 127-47 '99

**Philosophy, Political**

Larmore, Charles: The Moral Basis of Political Liberalism. 96 (12): 599-625 '99

**Physics, Philosophy of**

Callender, Craig: Reducing Thermodynamics to Statistical Mechanics: The Case of Entropy. 96 (7): 348-73 '99
Huggett, Nick: Atomic Metaphysics. 96 (1): 5-24 '99

**Preston, Beth**

Millikan, Ruth Garrett: Wings, Spoons, Pills, and Quills: A Pluralist Theory of Function. 96 (4): 191-206 '99

**Probability**

Christensen, David: Measuring Confirmation. 96 (9): 437-61 '99

**Property**

Yi, Byeong-uk: Is Two a Property? 96 (4): 163-90 '99

**Punishment**

Hershenov, David B.: Restitution and Revenge. 96 (2): 79-94 '99

**Putnam, Hilary**

Douven, Igor: Putnam's Model-theoretic Argument Reconstructed. 96 (9): 479-90 '99

**Quantum Mechanics**

Huggett, Nick: Atomic Metaphysics. 96 (1): 5-24 '99

**Quine, W. V.**

Sher, Gila: Is There a Place for Philosophy in Quine's Theory? 96 (10): 491-524 '99

**Rational Choice Theory**

Alexander, Jason: Bargaining with Neighbors: Is Justice Contagious? By Jason Alexander and Brian Skyrms. 96 (11): 588-98 '99

Skyrms, Brian: Bargaining with Neighbors: Is Justice Contagious? By Jason Alexander and Brian Skyrms. 96 (11): 588-98 '99

**Rawls, John**

Larmore, Charles: The Moral Basis of Political Liberalism. 96 (12): 599-625 '99

**Responsibility**

Kane, Robert: Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism. 96 (5): 217-40 '99

**Revenge**

Hershenov, David B.: Restitution and Revenge. 96 (2): 79-94 '99

**Science, Philosophy of**

Callender, Craig: Reducing Thermodynamics to Statistical Mechanics: The Case of Entropy. 96 (7): 348-73 '99

Christensen, David: Measuring Confirmation. 96 (9): 437-61 '99

Keeley, Brian L.: Of Conspiracy Theories. 96 (3): 109-26 '99

Lange, Marc: Calibration and the Epistemological Role of Bayesian Conditionalization. 96 (6): 294-324 '99

**Semantics**

Douven, Igor: Putnam's Model-theoretic Argument Reconstructed. 96 (9): 479-90 '99

Fodor, Jerry: All at Sea in Semantic Space: Churchland on Meaning Similarity. By Jerry Fodor and Ernie Lepore. 96 (8): 381-403 '99

Lepore, Ernie: All at Sea in Semantic Space: Churchland on Meaning Similarity. By Jerry Fodor and Ernie Lepore. 96 (8): 381-403 '99

Morreau, Michael: Supervaluation Can Leave Truth-value Gaps after All. 96 (3): 148-56 '99

Tiffany, Evan: Semantics San Diego Style. 96 (8): 416-29 '99

**Shapiro, Stewart**

Azzouni, Jody: Comments on Shapiro. 96 (10). 541-44 '99

Field, Hartry: Deflating the Conservativeness Argument. 96 (10): 533-40 '99

**Statistical Mechanics**

Callender, Craig: Reducing Thermodynamics to Statistical Mechanics: The Case of Entropy. 96 (7): 348-73 '99

**Supervaluation**

Morreau, Michael: Supervaluation Can Leave Truth-value Gaps after All. 96 (3): 148-56 '99

**Tarski, Alfred**

Azzouni, Jody: Comments on Shapiro. 96 (10): 541-44 '99
Field, Hartry. Deflating the Conservativeness Argument. 96 (10)· 533-40 '99
Williams, Michael: Meaning and Deflationary Truth. 96 (11): 545-64 '99

**Testimony**

Jones, Karen: Second-hand Moral Knowledge. 96 (2): 55-78 '99

**Trust**

Jones, Karen: Second-hand Moral Knowledge. 96 (2): 55-78 '99

**Truth**

Azzouni, Jody: Comments on Shapiro. 96 (10): 541-44 '99
Douven, Igor: Putnam's Model-theoretic Argument Reconstructed. 96 (9): 479-90 '99
Field, Hartry: Deflating the Conservativeness Argument. 96 (10). 533-40 '99
McCall, Storrs: Can a Turing Machine Know that the Gödel Sentence Is True? 96 (10): 525-32 '99
Sider, Theodore: Presentism and Ontological Commitment. 96 (7): 325-47 '99
Williams, Michael: Meaning and Deflationary Truth. 96 (11): 545-64 '99

**Turing Machines**

McCall, Storrs: Can a Turing Machine Know that the Gödel Sentence Is True? 96 (10)· 525-32 '99

**Vagueness**

Morreau, Michael: Supervaluation Can Leave Truth-value Gaps after All. 96 (3): 148-56 '99

**Van Fraassen, Bas C.**

Lange, Marc: Calibration and the Epistemological Role of Bayesian Conditionalization. 96 (6): 294-324 '99

**Vision**

Sorensen, Roy: Seeing Intersecting Eclipses. 96 (1): 25-49 '99

**Will**

Holton, Richard: Intention and Weakness of Will. 96 (5): 241-62 '99

## INDEX OF AUTHORS

Including Contributing Authors and Reviewers

**Alexander, Jason,** Bargaining with Neighbors: Is Justice Contagious? By Jason Alexander and Brian Skyrms. 96 (11): 588-98 '99

**Almog, Joseph,** Nothing, Something, Infinity. 96 (9): 462-78 '99

**Avigad, Jeremy,** Review of Solomon Feferman· *In the Light of Logic.* 96 (12): 638-42 '99

**Azzouni, Jody,** Comments on Shapiro. 96 (10): 541-44'99

**Blackburn, Simon,** Review of John Searle: *Mind, Language, and Society: Philosophy in the Real World.* 96 (12): 626-29 '99

**Callender, Craig,** Reducing Thermodynamics to Statistical Mechanics: The Case of Entropy. 96 (7): 348-73 '99

**Cavell, Marcia,** Review of Jonathan Lear: *Open Minded: Working Out the Logic of the Soul.* 96 (5): 263-69 '99

**Christman, John**, Review of Alfred R. Mele: *Autonomous Agents: From Self-control to Autonomy*. 96 (2): 95-99 '99

**Christensen, David**, Measuring Confirmation. 96 (9): 437-61 '99

**D'Arms, Justin**, Sex, Justice, and the Theory of Games 96 (12): 615-27 '99

**Dennett, Daniel C.**, Review of John Haugeland: *Having Thought: Essays in the Metaphysics of Mind*. 96 (8): 430-35 '99

**Douven, Igor**, Putnam's Model-theoretic Argument Reconstructed. 96 (9): 479-90 '99

**Field, Hartry**, Deflating the Conservativeness Argument. 96 (10): 533-40 '99

**Fodor, Jerry**, All at Sea in Semantic Space: Churchland on Meaning Similarity. By Jerry Fodor and Ernie Lepore. 96 (8): 381-403 '99

**Goldman, Alvin I.**, Internalism Exposed. 96 (6): 271-93 '99

**Hershenov, David B.**, Restitution and Revenge. 96 (2): 79-94 '99

**Hetherington, Stephen**, Knowing Failably. 96 (11): 565-87 '99

**Hintikka, Jaakko**, The Emperor's New Intuitions. 96 (3): 127-47 '99

**Holton, Richard**, Intention and Weakness of Will. 96 (5): 241-62 '99

**Horowitz, Gregg M.**, Review of Noel Carroll: *A Philosophy of Mass Art*. 96 (2): 99-105 '99

**Huggett, Nick**, Atomic Metaphysics. 96 (1): 5-24 '99

**Jones, Karen**, Second-hand Moral Knowledge. 96 (2): 55-78 '99

**Kane, Robert**, Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism. 96 (5): 217-40 '99

**Katz, Bernard D.**, On a Supposed Counterexample to Modus Ponens. 96 (8): 404-15 '99

**Keeley, Brian L.**, Of Conspiracy Theories. 96 (3): 109-26 '99

**Kelly, Michael**, Review of Anthony J. Cascardi: *Consequences of Enlightenment*. 96 (12): 630-37 '99

**Lange, Marc**, Calibration and the Epistemological Role of Bayesian Conditionalization. 96 (6): 294-324 '99

**Larmore, Charles**, The Moral Basis of Political Liberalism. 96 (12): 599-625 '99

**Lepore, Ernie**, All at Sea in Semantic Space: Churchland on Meaning Similarity. By Jerry Fodor and Ernie Lepore. 96 (8): 381-403 '99

**McCall, Storrs**, Can a Turing Machine Know that the Gödel Sentence Is True? 96 (10): 525-32 '99

**Millikan, Ruth Garrett**, Wings, Spoons, Pills, and Quills: A Pluralist Theory of Function. 96 (4): 191-206 '99

**Morreau, Michael**, Supervaluation Can Leave Truth-value Gaps after All. 96 (3): 148-56 '99

**Resnik, Michael D.**, Review of Jerrold J. Katz: *Realistic Rationalism*. 96 (4): 207-11 '99

**Sher, Gila**, Is There a Place for Philosophy in Quine's Theory? 96 (10): 491-524 '99

**Sider, Theodore**, Presentism and Ontological Commitment. 96 (7): 325-47 '99

**Skyrms, Brian**, Bargaining with Neighbors: Is Justice Contagious? By Jason Alexander and Brian Skyrms. 96 (11): 588-98 '99

**Sorensen, Roy**, Seeing Intersecting Eclipses. 96 (1): 25-49 '99

**Tait, W. W.,** Review of George Boolos: *The Logic of Provability.* **96** (1): 50-53 '99

**Tiffany, Evan,** Semantics San Diego Style. **96** (8): 416-29 '99

**Williams, Michael,** Meaning and Deflationary Truth. **96** (11): 545-64 '99

**Yi, Byeong-uk,** Is Two a Property? **96** (4): 163-90 '99

## INDEX OF BOOKS REVIEWED

Listed by Author of Item Reviewed

**Boolos, George:** *The Logic of Provability*—W. W. Tait. **96** (1): 50-53 '99
**Carroll, Noël:** *A Philosophy of Mass Art*—Gregg M. Horowitz. **96** (2): 99-105 '99
**Cascardi, Anthony J.:** *Consequences of Enlightenment*—Michael Kelly. **96** (12): 630-37 '99
**Feferman, Solomon:** *In the Light of Logic*—Jeremy Avigad. **96** (12): 638-42 '99
**Haugeland, John:** *Having Thought: Essays in the Metaphysics of Mind*—Daniel C. Dennett. **96** (8): 430-35 '99
**Katz, Jerrold J.:** *Realistic Rationalism*—Michael D. Resnik **96** (4): 207-11 '99
**Lear, Jonathan:** *Open Minded. Working Out the Logic of the Soul*—Marcia Cavell. **96** (5): 263-69 '99
**Mele, Alfred R.:** *Autonomous Agents: From Self-control to Autonomy*—John Christman. **96** (2): 95-99 '99
**Searle, John:** *Mind, Language and Society: Philosophy in the Real World*—Simon Blackburn. **96** (12): 626-29 '99